TRUSTMED-RL: LONG-HORIZON REINFORCEMENT LEARNING FOR EVIDENCE-GROUNDED CLINICAL DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical language models can produce correct diagnoses despite incomplete investigations and unsupported reasoning. To support long-horizon, evidence-grounded diagnosis, we introduce **TrustMed-RL**. Built from PubMed rare-disease cases and over 24,000 manually annotated image panels, it integrates interviews, examinations, testing, specialist consultation, and literature search through state-dependent actions. Our 8B vision–language policy, trained with clinically adapted GiGPO and coverage-adjusted diagnostic rewards, achieves 37.1% diagnostic accuracy on 2,500 evaluation cases, outperforming all evaluated open-weight baselines and improving over supervised fine-tuning by 12.4%. When success additionally requires acquiring at least 50% of supporting test evidence, TrustMed-RL achieves 32.5%, exceeding GPT-4o by 6.8%. Furthermore, it surpasses all evaluated baselines on MTMedDialog and multiple larger 27–32B models on AgentClinic. In physician review of 200 diagnostically accepted test-set trajectories, 83.0% receive evidential-grounding scores of 4–5 out of 5. Physicians' assessments suggest that these diagnostic trajectories are trustworthy and aligned with human diagnostic reasoning.

1 INTRODUCTION.

Clinical diagnosis requires an adaptive sequence of investigations. Clinicians elicit histories, select examinations and tests, integrate specialist interpretations, and revise the differential as evidence emerges (Balogh et al., 2015; Elstein & Schwarz, 2002). The ability to conduct such process becomes particularly important in rare, atypical, or serious disease, where the initial presentation may support several competing explanations. Recent medical agents make substantial progress toward interactive diagnosis. DoctorAgent-RL learns multi-turn inquiry policies, PulseMind optimizes multimodal consultation quality, and environments such as AgentClinic and MedEnv support sequential, tool-assisted investigation (Feng et al., 2026b; Xu et al., 2026a; Schmidgall et al., 2024; Fan et al., 2026). However, tool access and correct final answers alone do not establish that an agent acquired adequate supporting evidence. Models can still produce correct answers despite missing essential images and can fabricate findings to support their conclusions (Agbareia et al., 2025; von der Stück et al., 2026; Gu et al., 2025). Evaluating only the diagnostic endpoint can consequently conceal incomplete investigations, unsupported commitments, and failures to recognize unreliable evidence. These limitations motivate two questions: *Can training improve an agent's ability to coordinate the complete diagnostic pathway? Are its diagnostic gains accompanied by better evidence acquisition and more trustworthy decisions?* We introduce **TrustMed-RL**, an integrated agent framework, benchmark, and training protocol for *long-horizon, evidence-grounded diagnosis*. Our approach connects diagnostic outcomes to the observable clinical actions and evidence-acquisition processes that precede them. Our contributions are threefold:

1. **An end-to-end clinical agent framework.** We operationalize the diagnostic pathway through state-dependent actions spanning patient interviews, physical examination, test ordering, specialist consultation, literature search, and diagnosis or deferral. Encounters typically span 20–30 decisions, with a workup space covering 21 physical examinations, 42 diagnostic tests, and eight specialist departments. Specialist routing governs access to recorded interpretations, while explicit integrity judgments assess whether returned images

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

match the requested modality and anatomy. Action-level traces support failure localization, trajectory assessment, and targeted self-improvement data collection.

2. **A benchmark for groundedness and trustworthiness.** We extend RareArena with structured clinical records that separate patient-reported facts from clinician-acquired findings and incorporate more than 24,000 manually annotated panel-level images. Panel-specific evidence release limits unintended disclosure from composite figures, while ontology-based diagnosis matching supports consistent scoring. Evaluation jointly examines diagnostic correctness, evidence coverage, productive actions, responses to corrupted evidence, and capability-related deferral, making the evidential and procedural basis of diagnostic performance observable.
3. **Clinical adaptation of long-horizon policy learning.** Following supervised fine-tuning, we train the doctor policy using clinically adapted Group-in-Group Policy Optimization (GiGPO) and a ScalingInter-RL-based interaction curriculum (Feng et al., 2026a; Xi et al., 2025). To our knowledge, this is the first study to combine these methods for long-horizon multimodal clinical diagnosis. Our adaptations introduce clinical anchors for step-level comparisons, coverage-adjusted diagnostic rewards, local invalid-action penalties, and a curriculum coupling progressive interaction budgets with resource-aware case admission and outcome-adaptive sampling.

Experimental results show that the 8B TrustMed-RL policy achieves 37.1% diagnostic accuracy on 2,500 evaluation cases, leading all evaluated open-weight models, with findings coverage reaching 73.2% and productive actions achieving 8.5 per epsiode. When success additionally requires acquiring at least 50% of available test findings, TrustMed-RL achieves 32.5%, exceeding GPT-4o by 6.8 percentage points. Furthermore, Additional out-of-distribution experiments show both TrustMed-RL variants demonstrate strong generalization across three benchmarks. Source code and data are available at https://github.com/Trustmed-RL/TrustMed-RL

2 Related Work

Interactive clinical consultation. Recent medical agents extend static question answering toward interactive diagnosis. AgentClinic supports multimodal consultations and diagnostic tools, including evaluation of clinical-information coverage (Schmidgall et al., 2024). AMIE learns diagnostic dialogue through simulated consultations, while its multimodal extension introduces state-aware dialogue phases and tracks information gaps (Tu et al., 2025; Saab et al., 2026). PulseMind combines multimodal consultation data with comparison-based reinforcement learning and evaluates proactivity, diagnostic accuracy, usefulness, and language quality (Xu et al., 2026a). These studies establish interactive information gathering and multidimensional consultation assessment as central components of medical-agent evaluation.

Learning and coordinating diagnostic actions. DoctorAgent-RL applies GRPO to multi-turn doctor–patient interactions, learning questioning and diagnostic policies under varying dialogue budgets of two to ten turns (Feng et al., 2026b). MedEnv integrates interviewing, retrieval, test ordering, and chest-radiograph analysis, combining confidence-based process supervision with diagnostic outcome rewards (Fan et al., 2026). DxDirector guides the diagnostic process through requests for physician-assisted clinical operations and evaluates request usefulness, physician involvement, and diagnostic accountability (Xu et al., 2026b). These approaches advance policy learning and clinical-action coordination while demonstrating the importance of evaluating intermediate decisions alongside final diagnoses.

Our contribution. We build on these advances through the joint design of an executable clinical environment and an adaptive learning protocol for *grounded and trustworthy diagnosis*. Our environment connects interviewing, examination, testing, specialist consultation, and literature search to diagnosis or deferral through state-dependent actions. Findings are grounded in case records, specialist-department routing controls access to recorded interpretations, and scored integrity verdicts assess recognition of modality and anatomical mismatches. These controls make evidence acquisition, routing, integrity assessment, and diagnostic commitment explicitly observable and jointly scorable. We further adapt GiGPO through clinical anchors for step-level comparisons, coverage-adjusted diagnostic rewards, local invalid-action penalties, and an outcome-adaptive interaction curriculum. The resulting framework links policy improvement to observable evidence-acquisition

behavior and supports analysis of whether diagnostic success is accompanied by sufficient evidence access, appropriate integrity judgments, and reliable clinical actions.

3 PRELIMINARIES

Formulation Following formulations of multi-turn LLM agents Xi et al. (2025; 2026); Feng et al. (2026a), We model a clinical consultation as a finite-horizon partially observable Markov decision process (POMDP) Kaelbling et al. (1998), as illustrated in Figure 1 in Appendix B. A case $x \sim \mathcal{D}$ induces the POMDP $M_x = \langle S, O, A, P, \Omega, r, \rho_0^x, H \rangle$. where S, O, and A are the state, observation, and textual action spaces; P and Ω are the transition and observation kernels; r is the reward function; ρ_0^x is the case-conditioned initial-state distribution; and H bounds the number of policy decisions.

The latent state $s_t = (z_x, m_t, \ell_t, \varphi_t, b_t, \zeta_t)$ comprises case information z_x, interaction history m_t, evidence state ℓ_t, the current interaction phase φ_t (history taking, diagnostic workup, evidence-integrity checking, or conclusion), remaining interaction budgets b_t, and integrity state ζ_t. At turn t, the doctor agent receives $o_t \sim \Omega(\cdot \mid s_t)$, containing released clinical information and the current action menu; unreleased findings, the reference diagnosis, and ground-truth integrity labels remain hidden. Given the policy context h_t serialized from the observable interaction history, the doctor agent acts according to

$$a_t \sim \pi_\theta(\cdot \mid h_t), \qquad s_{t+1} \sim P(\cdot \mid s_t, a_t), \qquad r_t = r(s_t, a_t, s_{t+1}). \tag{1}$$

The environment parses and validates a_t against the admissible set $A(s_t) \subseteq A$, determined by the current phase, remaining budgets, and evidence and integrity states. The patient simulator and clinical tools are environment components; only the doctor agent policy is optimized. The policy context h_t is reconstructed at each turn from the observable interaction history, including released findings and previous doctor actions while hidden case information is consistently excluded. A consultation produces a trajectory $\tau = (o_0, a_0, \ldots, o_{T-1}, a_{T-1}, o_T)$, with $T \leq H$, and terminates upon diagnosis or abstention, or is terminated when an interaction limit is reached.

4 AGENT FRAMEWORK

State-dependent clinical interaction. Our agent harness implements an end-to-end diagnostic encounter, integrating history-taking, physical examination, diagnostic testing, specialist consultation, and medical information search. Each episode follows a real published case and typically spans 20–30 decisions. The workup action space includes **21 physical examinations**, **42 diagnostic tests** and consultation across **eight specialist departments**. The agent selects actions according to the consultation phase, available evidence, and remaining interaction budget. Each action changes the environment state and subsequent decision options, requiring coordination of evidence acquisition, specialist routing, and diagnostic reasoning before diagnosis or human deferral.

Observable evidence and decision reliability. The harness makes evidence acquisition and release explicit: ordering a study and obtaining its specialist interpretation are distinct operations, report release requires appropriate routing, and unavailable examinations return an explicit availability status. Invalid actions receive deterministic feedback and remain recorded in the trajectory. These controls support joint assessment of diagnostic correctness, evidence coverage, routing, responses to corrupted evidence, action validity, and capability-related deferral. A diagnosis can therefore be assessed alongside the evidence available when it was made and the reliability of the actions leading to it, exposing unsupported diagnostic commitments and intermediate failures that final-answer accuracy alone may overlook.

Failure analysis and self-improvement. Every action is linked to its preceding state, returned observation, and evaluation feedback. This trace supports localization of failures to specific decisions, including missed evidence, incorrect specialist routing, invalid requests, and inappropriate diagnostic commitment or deferral. It enables systematic examination of how a consultation failed and which observable behaviors undermine trustworthiness. The same structured feedback supports trajectory-quality assessment, selection of informative successes and failures, and construction of targeted self-improvement data across the complete diagnostic pathway.

5 TRAINING FRAMEWORK

We introduce a unified training framework for a doctor agent operating in a long-horizon, multi-state, multi-action clinical environment. The policy must coordinate patient questioning, evidence acquisition, report review, and diagnosis as the consultation state evolves. We adapt Group-in-Group Policy Optimization (GiGPO) (Feng et al., 2026a), with Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as an episode-level alternative, and extend the interaction curriculum of ScalingInter-RL (Xi et al., 2026) within a shared clinical training protocol. Our adaptations connect clinical state comparisons, delayed diagnostic outcomes, action validity, and case-dependent interaction budgets to an explicit objective: learning to reach correct diagnoses after acquiring supporting evidence.

As illustrated in Appendix Figure 2, Stage I performs supervised fine-tuning (SFT) on curated multi-turn consultations generated by GPT-5-mini and Qwen3.8-max. Stage II initializes from the SFT policy and optimizes the doctor agent through simulated clinical interaction. This two-stage process establishes consultation behavior before refining the complete diagnostic pathway through reinforcement learning.

Evidence-conditioned diagnosis reward. Let $d_i \in [0, 1]$ denote diagnostic correctness, M_i the number of image-bearing evidence records available for case i, and L_i the number of distinct records whose specialist reports have been released during consultation. For $M_i > 0$, we define coverage as $c_i = \min\{1, L_i/M_i\}$ and the diagnosis reward as $R_i^{\mathrm{dx}} = d_i(0.10 + 0.90\sqrt{c_i})$. For $M_i = 0$, we leave diagnosis credit unchanged, $R_i^{\mathrm{dx}} = d_i$. This reward couples diagnostic correctness with evidence availability: coverage alone earns no diagnosis credit, and, on image-bearing cases, a correct diagnosis without released reports receives only the base credit. Ordering or viewing an image does not count toward coverage, and repeated requests do not increase L_i. The adjustment is applied once to terminal diagnosis credit; intermediate evidence acquisition receives no separate positive reward.

Delayed outcomes and local action penalties. Let R_i denote the terminal task reward, using R_i^{dx} for diagnosis outcomes, and let T_i be the consultation length. Set $v_{it} = 0$ for a penalized format violation and $v_{it} = 1$ otherwise. The scores supplied to the episode- and step-level credit estimators are $q_{it}^E = R_i - 0.01(1 - v_{it})$ and $q_{it}^S = 0.99^{T_i - t} R_i - 0.01(1 - v_{it})$, respectively. Discounting propagates delayed diagnostic outcomes to earlier consultation decisions while favoring efficient successful resolution. Across action types, each format violation incurs the same fixed local cost of 0.01, independent of realized diagnosis credit. Thus, terminal outcomes supervise the diagnostic pathway, while malformed intermediate actions receive immediate, localized penalties that are neither discounted nor propagated to other turns.

Credit assignment across clinical states. For each case, we collect $K \geq 2$ consultations with matched initial context, interaction budgets, and corruption realization. Each rollout contributes one reference score to episode-level statistics, preventing longer consultations from dominating the baseline. For GiGPO's step-level credit, an environment-defined abstraction $\tilde{s}_{it} = \phi(s_{it})$ groups decisions by consultation phase, progress, and evidence status. These clinical anchors align decisions at comparable stages of evidence acquisition despite different textual histories and consultation lengths. GRPO uses the same clinical rewards with episode-level credit only. Step-level credit is zero when an anchor lacks sufficient comparisons, and a variance floor stabilizes normalization in low-variance groups.

Selective reinforcement of correct diagnoses. Let A_{it}^E and A_{it}^S denote the resulting episode- and step-level advantages. We strengthen eligible positive updates after normalization:

$$A_{it} = \left[1 + (\alpha_{\mathrm{dx}} - 1)e_{it}\right] A_{it}^E + \omega A_{it}^S, \tag{2}$$

$$e_{it} = \chi_i v_{it}\, \mathbf{1}\{A_{it}^E > 0\}\, \mathbf{1}\{A_{it}^E + \omega A_{it}^S > 0\}. \tag{3}$$

Here $\alpha_{\mathrm{dx}} \geq 1$, $\omega \geq 0$, and $\chi_i = 1$ only for a fully correct diagnosis on a case. Setting $\omega = 0$ yields episode-level credit. Applying the multiplier after normalization strengthens valid, already-favorable actions in correct consultations without rescaling other group members' advantages or reversing a nonpositive combined update. Evidence acquisition affects eligibility indirectly through the coverage-adjusted reward.

Outcome-adaptive task sampling. We prioritize tasks using variation in observed outcomes under the current policy. At each visit to task x, rollout scores account for terminal reward and cumulative invalid-action costs:

$$\widetilde{R}_{xi} = \mathrm{clip}_{[0,1]}\left(\frac{R_{xi} - \lambda_{\mathrm{inv}} \sum_{t=1}^{T_{xi}} (1 - v_{xit})}{R^{\mathrm{ref}}_{k_x}(n)}\right), \qquad \sigma_x = \mathrm{Std}^{K}_{i=1}(\widetilde{R}_{xi}), \tag{4}$$

where $R^{\mathrm{ref}}_{k_x}(n) > 0$ is the task-type reference reward and Std denotes population standard deviation. Sampling priorities track a moving average of visit dispersion, with offline difficulty providing the initial prior. Following prioritized level replay (Jiang et al., 2021), we combine priority ranks with staleness within task-type strata, reserve uniform exploration slots, and retire consistently high-performing, low-dispersion tasks.

A clinical interaction curriculum. We extend ScalingInter-RL by coupling progressive interaction budgets with case admission and outcome-adaptive sampling. At update n, phase $p(n)$ determines the budget ceiling and admits tasks according to their estimated clinical interaction requirements:

$$\mathcal{D}_n = \{x : h_{x,J_x(n)} \le p(n)\}, \qquad \mathbf{b}_1 \preceq \cdots \preceq \mathbf{b}_P = \mathbf{b}_{\mathrm{eval}}. \tag{5}$$

Here $J_x(n)$ counts completed visits, $h_{x,J_x(n)}$ is the estimated admission phase, and $\mathbf{b}_p$ specifies the interaction-budget ceilings. The curriculum progressively introduces cases requiring longer consultations, while adaptive sampling identifies informative cases within the admitted pool. Matched budgets within each rollout group preserve comparable training conditions. Together with evidence-conditioned rewards and local action penalties, these adaptations make grounded diagnosis and reliable consultation behavior explicit training objectives for the doctor agent. Curriculum details and hyperparameters appear in Appendices B.2 and B.3.

6 Groundedness and Trustworthiness

Groundedness and trustworthiness are explicit design requirements across our data annotation, preprocessing, agent harness, and training framework. Long-horizon clinical diagnosis requires an agent to acquire evidence, navigate evolving clinical states, and coordinate heterogeneous actions before committing to a diagnosis. We therefore design these components as an integrated system that makes evidence provenance, information access, intermediate decisions, and diagnostic outcomes observable throughout the consultation. Their interdependence supports both faithful evaluation and systematic policy improvement.

Traceable clinical evidence through manual annotation. We extend RareArena (Chen et al., 2026) into a multimodal, long-horizon agentic benchmark with structured clinical evidence. Our image preprocessing pipeline draws on 44,277 registered case figures from PubMed Central. Large-scale manual annotation identifies semantic panels and records their identity, source order, modality, anatomic site, laterality, and pixel coordinates, yielding more than **24,000 panel-level crops**. Each image crop is linked to its source examination, allowing evidence to be released selectively by examination and routed to the appropriate specialist. This mapping makes the visual evidence presented to the agent traceable while limiting inadvertent disclosure of unrelated findings within composite figures.

Atomic clinical records and controlled interaction. Our preprocessing pipeline organizes cases into **OSEC**-style atomic clinical facts, separating patient-reported information from clinician-discovered findings (Figure 9). This distinction defines which facts the simulated patient may report and which findings require further clinical investigation. The agent harness connects these records to controlled evidence release and tracks acquisition across the consultation. Together, atomic facts and panel-specific access make it possible to determine whether an action reveals new evidence, repeats an earlier request, or leaves available findings unexplored. This directly supports evaluation of action productivity and evidence coverage across variable-length, multi-state consultations.

Ontology-grounded diagnostic supervision. Reliable policy learning requires diagnostic feedback that remains consistent across terminology variants. Integrating MedGen, MONDO, and ORDO, our ontology-based matching framework maps predicted diagnoses and reference labels

to standardized disease concepts. Compared with lexical-overlap scoring and LLM-only diagnostic assessment, our framework incoportaes ontology-based normalization with LLM judgment. This combination recognizes clinically equivalent diagnoses despite lexical variation and assesses semantic similarity against standardized disease concepts, preserving clinically meaningful distinctions between related diagnoses. The same scoring procedure underpins training rewards and evaluation, providing a consistent criterion for diagnostic correctness. Table 12 presents reference diagnosis pairs illustrating accepted equivalences and distinctions, making the scoring criteria transparent and auditable.

Retrievable medical knowledge at scale. To support reasoning beyond the agent's parametric knowledge, we construct a medical corpus following MedRAG (Zhao et al., 2025), comprising 29,650,414 passages. These include 29,166,067 PubMed title-plus-abstract passages from the 2026 MEDLINE baseline and daily updates through July 30, 2026; 358,500 StatPearls paragraph-level passages titled by section paths; and 125,847 passages from 18 textbooks. All passages are embedded using the MedCPT article encoder. Retrieval provides identifiable external references that complement case-specific observations and support inspection of the knowledge available during diagnostic reasoning.

Evidence-aware training for long-horizon diagnosis. We adapt GiGPO (Feng et al., 2026a), with GRPO (Shao et al., 2024) as an episode-level alternative, and extend ScalingInter-RL (Xi et al., 2026) within a common clinical training protocol. Clinical anchors support credit assignment across different consultation histories and evidence states, while local format penalties discourage invalid intermediate actions. Outcome-adaptive sampling and progressively expanded interaction budgets expose the policy to cases with increasingly demanding interaction requirements. Crucially, diagnostic rewards depend on both correctness and released-report coverage (Evidence-conditioned diagnosis reward), and selective reinforcement strengthens eligible positive updates from correct consultations. These adaptations make evidence acquisition and reliable interaction explicit training objectives across the complete diagnostic pathway. Coverage operationalizes evidence acquisition as a grounding proxy; it does not itself verify faithful use of the acquired evidence.

Auditable evaluation and iterative self-improvement. We evaluate complete consultations through diagnostic correctness, evidence coverage, action productivity, and performance across coverage levels. Joint assessment distinguishes successful evidence acquisition from correct terminal prediction and reveals behaviors such as repeated requests, incomplete investigation, and correct diagnoses reached with limited evidence coverage. These observations provide actionable feedback for subsequent task sampling, policy training, and evaluation, supporting **iterative self-improvement** under consistent criteria. Each component supplies a prerequisite for this feedback loop: annotation establishes provenance, preprocessing defines observable facts, the harness records evidence access, ontology-based scoring provides consistent outcome feedback, and training converts that feedback into policy updates. Their integration makes long-horizon diagnostic behavior auditable and grounds assessments of robustness and trustworthiness in observable actions, acquired evidence, and outcomes.

7 EXPERIMENT

Datasets and implementation. We train Qwen3-VL-8B-Instruct (Bai et al., 2025) on RareArena (Chen et al., 2026) and evaluate out-of-distribution generalization on AgentClinic-NEJM, AgentClinic-MedQA (Schmidgall et al., 2024), and MTMedDialog (Feng et al., 2026b). We use `gpt-5.6-terra` to convert RareArena case reports and associated PMC figures into diagnosis-blind OSCE-style cases, separating patient-reported histories from examination findings and diagnostic evidence available through workup. Following TruthRL (Wei et al., 2025), hierarchical diagnostic probes classify cases as easy, medium, hard, or empirically unsolved under the tested models and budgets. We evaluate GPT-5-mini as a diagnostic judge against ontology-derived reference pairs and physician-consensus labels, and use GPT-4.1-mini for online reward computation and GPT-5.1-mini for evaluation (Appendix B). Sparse outcome rewards cover diagnosis, corrupted-evidence detection, and appropriate deferral (Appendix C).

7.1 DIAGNOSIS EVALUATION

Condition extraction and semantic matching. For case i, let $\mathcal{G}_i$ and $\mathcal{P}_i$ denote the ground-truth and predicted condition sets, with $m_i = |\mathcal{G}_i|$ and $n_i = |\mathcal{P}_i|$. Each judge $j \in \{1, \dots, J\}$ labels candidate pairs (g, p) as FULL, CORE, PARTIAL, or NULL (Appendix E.1). The corresponding weights satisfy $w_{\mathrm{F}} = 1 > w_{\mathrm{C}} > 0.5 \geq w_{\mathrm{P}} > w_{\varnothing} = 0$.

To mitigate judge sensitivity, a deterministic offline oracle based on MedGen, MONDO, and ORDO supplies a label floor $b_i(g, p) \in \{\varnothing, \mathrm{C}, \mathrm{F}\}$. Reward computation uses the augmented label $\widetilde{\ell}_{ij}(g, p) = \max_{\prec}\{\ell_{ij}(g, p), b_i(g, p)\}$, where $\varnothing \prec \mathrm{P} \prec \mathrm{C} \prec \mathrm{F}$.

Soft diagnostic scores. Let F_{ij}, C_{ij}, and P_{ij} count the FULL, CORE, and PARTIAL matches in a one-to-one matching. The soft matched mass is $s_{ij} = w_{\mathrm{F}}F_{ij} + w_{\mathrm{C}}C_{ij} + w_{\mathrm{P}}P_{ij}$, with $0 \leq s_{ij} \leq \min\{m_i, n_i\}$. The residual false-positive and false-negative masses are $\mathrm{FP}^s_{ij} = n_i - s_{ij}$ and $\mathrm{FN}^s_{ij} = m_i - s_{ij}$. For $\alpha, \beta \geq 0$, the soft Tversky score (Salehi et al., 2017) is

$$T^{(ij)}_{\alpha,\beta} = \frac{s_{ij}}{s_{ij} + \alpha(n_i - s_{ij}) + \beta(m_i - s_{ij})} = \frac{s_{ij}}{\alpha n_i + \beta m_i}, \qquad \alpha + \beta = 1. \tag{6}$$

Here, α and β control the penalties for unsupported predicted diagnoses and missed ground-truth diagnoses, respectively. All reported settings satisfy $\alpha + \beta = 1$. For $m_i, n_i \geq 1$, soft precision and recall are $\mathrm{Prec}_{ij} = s_{ij}/n_i$ and $\mathrm{Rec}_{ij} = s_{ij}/m_i$, measuring semantic support for predicted conditions and coverage of ground-truth conditions. When $s_{ij} > 0$, the Tversky score is their weighted harmonic mean: $(T^{(ij)}_{\alpha,\beta})^{-1} = \alpha/\mathrm{Prec}_{ij} + \beta/\mathrm{Rec}_{ij}$.

Diagnostic accuracy. Accuracy requires every ground-truth condition to receive a FULL or CORE match, $\mathrm{Acc}_{ij} = \mathbf{1}\{F_{ij} + C_{ij} = m_i\}$; strict accuracy requires FULL matches throughout, $\mathrm{Acc}^{\mathrm{strict}}_{ij} = \mathbf{1}\{F_{ij} = m_i\}$. Additional predicted conditions are penalized by the soft scores but do not invalidate these accuracy criteria. Non-diagnosis outcomes receive zero.

Evidence acquisition. Let $\mathcal{T}_i$ contain the available laboratory, physiologic, imaging, and pathology test records, and let $\mathcal{R}_i \subseteq \mathcal{T}_i$ contain those whose nonempty findings reach the doctor agent before termination. Findings coverage is $c_i = |\mathcal{R}_i|/|\mathcal{T}_i|$ when $|\mathcal{T}_i| > 0$, and $c_i = 0$ otherwise. To assess diagnostic accuracy jointly with evidence acquisition, we report

$$\mathrm{Acc@50} = \frac{100}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}\mathrm{Acc}_{ij}\,\mathbf{1}\{c_i \geq 0.5\}, \quad \mathrm{Strict@50} = \frac{100}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}\mathrm{Acc}^{\mathrm{strict}}_{ij}\,\mathbf{1}\{c_i \geq 0.5\}. \tag{7}$$

Both metrics require at least 50% findings coverage and retain all N cases in the denominator; cases without available test records contribute zero.

7.2 EXPERIMENTAL RESULTS

We evaluate diagnostic performance on 2,500 clean cases, its retention under an explicit evidence-acquisition requirement, and the ability to conduct a multistage diagnostic workflow in an interactive clinical environment.

Leading diagnostic performance among open-weight models. TrustMed-RL (GiGPO, 8B) achieves the highest open-weight accuracy (37.1%), precision (31.8%), and recall (34.6%) (Table 1). Its accuracy exceeds the 8B SFT model by 12.4 percentage points, an approximately 50% relative gain, and HealthGPT-Pro-27B, the strongest external open-weight baseline by accuracy, by 2.6 points. GiGPO also improves over GRPO (34.5%) and approaches GPT-4o (37.9%). GPT-5 remains the strongest model overall.

Diagnostic gains remain robust under evidence requirements. GiGPO achieves the highest open-weight Acc@50 (32.5%) and Strict@50 (13.5%), showing that its advantage persists when diagnostic correctness must be accompanied by substantial findings acquisition. Its Acc@50 exceeds GPT-4o by 6.8 percentage points and HealthGPT-Pro-27B by 9.0 points. Moreover, GiGPO and GRPO incur the two smallest relative accuracy reductions across all evaluated models: 12.3% and

Table 1: Diagnostic performance on 2,500 clean cases per evaluation run. All scores are percentages. SFT, GiGPO, and GRPO each use three training seeds, with three separately seeded evaluation passes per trained model (nine runs per method). Where reported, means and sample standard deviations (std) are computed across the three evaluation passes. $\Delta_{\mathrm{Acc}} = 100(\mathrm{Acc} - \mathrm{Acc@50})/\mathrm{Acc}$ and $\Delta_{\mathrm{strict}} = 100(\mathrm{Acc}^{\mathrm{strict}} - \mathrm{Strict@50})/\mathrm{Acc}^{\mathrm{strict}}$ measure relative accuracy reductions under the evidence-coverage requirement (lower is better). Non-diagnosis outcomes receive zero. Bold indicates the best result among open-weight models.

Model	Acc	$\mathrm{Acc}^{\mathrm{strict}}$	Prec.	Rec.	Acc@50	Strict@50	Δ_{Acc} (%)	Δ_{strict} (%)
Proprietary (closed-source) models								
GPT-5	54.2	22.0	42.9	50.0	43.2	17.7	20.3	19.5
GPT-4o	37.9	20.8	34.9	36.0	25.7	14.3	32.3	31.3
Open-weight models								
TrustMed-RL (GiGPO, 8B)	$\mathbf{37.1}_{\pm 1.1}$	$15.3_{\pm 1.0}$	$\mathbf{31.8}_{\pm 1.1}$	$\mathbf{34.6}_{\pm 1.1}$	$\mathbf{32.5}_{\pm 1.3}$	$\mathbf{13.5}_{\pm 0.6}$	12.3	11.9
TrustMed-RL (GRPO, 8B)	$34.5_{\pm 1.6}$	$13.6_{\pm 1.9}$	$28.9_{\pm 1.9}$	$32.1_{\pm 1.5}$	$30.8_{\pm 1.3}$	$12.2_{\pm 1.6}$	**10.7**	**10.9**
TrustMed-SFT (8B)	$24.7_{\pm 0.6}$	$9.0_{\pm 0.3}$	$19.0_{\pm 0.4}$	$22.9_{\pm 0.6}$	$18.8_{\pm 0.4}$	$6.9_{\pm 0.3}$	23.9	23.2
TrustMed-SFT (2B)	$13.0_{\pm 0.1}$	$3.8_{\pm 0.3}$	$9.3_{\pm 0.2}$	$12.3_{\pm 0.2}$	$8.9_{\pm 0.1}$	$2.7_{\pm 0.4}$	31.1	30.3
Qwen3-VL-8B-Instruct (Bai et al., 2025)	$16.3_{\pm 0.3}$	$6.4_{\pm 0.5}$	$13.7_{\pm 0.2}$	$15.3_{\pm 0.3}$	$6.6_{\pm 0.2}$	$2.7_{\pm 0.1}$	59.2	58.0
PulseMind-72B	$21.8_{\pm 0.3}$	$10.1_{\pm 0.3}$	$19.2_{\pm 0.2}$	$20.9_{\pm 0.1}$	$11.1_{\pm 0.1}$	$5.5_{\pm 0.3}$	49.4	45.8
HealthGPT-Pro-27B	$34.5_{\pm 0.7}$	$\mathbf{18.1}_{\pm 0.9}$	$30.8_{\pm 1.0}$	$32.6_{\pm 0.8}$	$23.5_{\pm 0.8}$	$12.7_{\pm 0.8}$	31.8	30.0
Qwen3-VL-32B-Instruct	30.4	14.9	26.9	28.6	20.9	10.2	31.2	31.4
MedGemma-27B-it (Sellergren et al., 2025)	30.4	14.9	27.6	28.7	20.9	10.7	31.1	28.4
Hulu-Med-32B (Jiang et al., 2025)	23.4	10.3	19.8	22.1	14.4	6.7	38.4	35.3
Lingshu-32B (Xu et al., 2026c)	21.5	10.0	19.5	20.6	13.9	7.0	35.5	29.3
QoQ-Med-VL-32B Dai et al. (2026)	13.4	4.8	11.5	13.1	8.6	3.7	36.3	21.8

10.7%, respectively, compared with 23.9% for SFT, 31.8% for HealthGPT-Pro-27B, and 32.3% for GPT-4o. GiGPO therefore retains 87.7% of its diagnostic accuracy when substantial findings acquisition is required These results show that the diagnostic improvements are not confined to episodes with insufficient evidence acquisition. Complementary physician assessment of findings interpretation, evidence use, and evidential grounding further supports the quality of the diagnostic reasoning (Appendix G).

RL supports sustained and productive agentic investigation. The diagnostic gains accompany stronger evidence acquisition and productive interaction (Table 5). GiGPO and GRPO achieve ordering coverage of 83.9% and 84.7% and findings coverage of 73.2% and 73.9%, respectively, exceeding GPT-5 (72.4% and 62.8%) and GPT-4o (62.0% and 53.5%). GiGPO's search use rises from 53.2% to 93.0% of episodes, its productive actions increase from 7.7 per episode under SFT to 8.5 for GiGPO and 9.1 for GRPO, surpassing GPT-4o (6.9) and matching GPT-5 (8.5). Since productive actions reveal previously unseen information, these results indicate that sustained interactions yield additional evidence. Together with the coverage-gated diagnostic results, these behaviors demonstrate learned agentic workup in long-horizon clinical interactions, with the policy coordinating evidence acquisition, specialist consultation, and search toward diagnostic resolution.

Out-of-distribution performance. We further evaluate interactive diagnostic performance on MTMedDialog, AgentClinic-MedQA, and AgentClinic-NEJM (Feng et al., 2026b; Schmidgall et al., 2024). Across all three benchmarks, models operate through multi-turn clinical interactions and produce free-form diagnoses rather than selecting from predefined answer options. As shown in Appendix Table D.1, both TrustMed-RL variants outperform all evaluated 7–8B baselines across the three datasets. On MedQA and NEJM, they achieve performance comparable to several substantially larger models and surpass multiple 27–32B baselines. On MTMedDialog, GiGPO and GRPO achieve the two highest accuracies among all evaluated models, at 48.7% and 49.5%, respectively. These results demonstrate competitive OOD performance and demonstrate learned diagnostic capabilities in distinct long-horizon, multi-turn clinical environments.

7.3 Ablation study

SFT and RL gains. Table 7 in Appendix D.2 shows that SFT improves solvable ACC by 12.30 and 9.49 percentage points for 2B and 8B, respectively. GiGPO and GRPO further improve 8B SFT by 13.84 and 12.09 points. Removing priority or adaptive sampling reduces solvable Strict by 2.69

Table 2: Agentic behavior for models in Table 1. Order and Findings measure mean test-record ordering and findings coverage over 2,452 eligible clean cases. Findings includes laboratory, physiologic, imaging, and pathology reports, excluding raw images. Prod. counts actions yielding new history facts, images, or findings; Workup counts workup turns, both per clean episode. Search measures clean episodes with any search; Invalid measures invalid clean-episode turns. Modality and Anatomy Catch measure exposure-conditioned mismatch detection. False Alarm is the false-positive rate on matched clean-image gate-replay controls. All metrics except Prod. and Workup are percentages; — denotes unavailable results.

Model	Order	Findings	Prod.	Workup	Search	Invalid	Modality Catch	Anatomy Catch	False Alarm
Proprietary (closed-source) models									
GPT-5	72.4	62.8	8.5	12.8	63.2	0.0	97.8	88.7	4.8
GPT-4o	62.0	53.5	6.9	10.9	75.6	0.0	95.4	85.7	6.4
Open-weight models									
TrustMed-RL (GiGPO, 8B)	$83.9_{\pm 4.0}$	$73.2_{\pm 3.9}$	$8.5_{\pm 0.4}$	$17.8_{\pm 0.8}$	$93.0_{\pm 2.5}$	$4.0_{\pm 1.1}$	$82.2_{\pm 1.1}$	$52.6_{\pm 4.2}$	$2.2_{\pm 0.8}$
TrustMed-RL (GRPO, 8B)	$\mathbf{84.7}_{\pm 1.3}$	$\mathbf{73.9}_{\pm 1.1}$	$\mathbf{9.1}_{\pm 0.2}$	$18.3_{\pm 0.6}$	$89.3_{\pm 4.7}$	$2.6_{\pm 0.2}$	$85.4_{\pm 7.9}$	$60.6_{\pm 2.4}$	$2.4_{\pm 0.5}$
TrustMed-SFT (8B)	$66.4_{\pm 0.8}$	$57.1_{\pm 0.7}$	$7.7_{\pm 0.0}$	$10.2_{\pm 0.1}$	$53.2_{\pm 1.4}$	$\mathbf{0.6}_{\pm 0.0}$	$92.2_{\pm 1.9}$	$\mathbf{73.3}_{\pm 1.4}$	$3.9_{\pm 0.8}$
TrustMed-SFT (2B)	$59.4_{\pm 0.1}$	$51.3_{\pm 0.2}$	$7.2_{\pm 0.0}$	$11.1_{\pm 0.1}$	$50.7_{\pm 1.5}$	$6.9_{\pm 0.1}$	$80.8_{\pm 5.7}$	$36.2_{\pm 10.1}$	$5.0_{\pm 0.4}$
Qwen3-VL-8B-Instruct	$36.1_{\pm 1.0}$	$30.8_{\pm 0.8}$	$4.7_{\pm 0.0}$	$5.6_{\pm 0.1}$	$18.1_{\pm 0.9}$	$7.1_{\pm 0.3}$	$\mathbf{92.9}_{\pm 5.0}$	$57.0_{\pm 3.6}$	$4.6_{\pm 2.3}$
PulseMind-72B	$45.4_{\pm 0.6}$	$37.6_{\pm 0.6}$	$5.1_{\pm 0.0}$	$7.2_{\pm 0.0}$	$36.6_{\pm 1.7}$	$5.3_{\pm 0.0}$	—	—	—
HealthGPT-Pro-27B	$61.6_{\pm 0.3}$	$53.3_{\pm 0.3}$	$6.3_{\pm 0.0}$	$12.3_{\pm 0.7}$	$53.2_{\pm 4.5}$	$17.0_{\pm 3.8}$	$82.0_{\pm 3.4}$	$70.9_{\pm 0.7}$	$5.9_{\pm 0.6}$
Qwen3-VL-32B-Instruct	55.0	47.1	6.0	11.5	61.3	5.2	83.6	68.8	3.9
MedGemma-27B-it	61.2	52.9	6.7	18.1	77.8	41.0	42.9	61.1	7.2
Hulu-Med-32B	49.9	42.6	5.9	11.8	51.4	15.4	51.2	40.8	2.7
Lingshu-32B	52.5	45.2	6.0	10.3	43.2	7.4	48.6	56.2	4.7
QoQ-Med-VL-32B	53.6	46.8	5.8	12.6	80.4	10.3	36.4	37.3	3.1

or 1.82 points. Jointly removing adaptive sampling and progressive horizons produces the largest Strict decline, from 18.20% to 14.43%, and lowers Strict@3 by 5.27 points.

Training behavior. Late-training statistics (Table 8) reveal broader deficits under joint removal: diagnostic and hard-band success fall by 6.17 and 8.19 percentage points, while corruption-task success and exposed-mismatch detection fall by 13.80 and 12.93 points, the largest integrity-check losses. These deficits persist despite higher evidence coverage (69.50% versus 66.32%) and more turns per episode (23.80 versus 21.81). Individually removing priority or adaptive sampling also yields substantial diagnostic deficits (7.18 and 9.48 points), despite smaller held-out $\mathrm{Acc}^{\mathrm{strict}}$ losses. GRPO nearly matches terminal reward, yet trails diagnostic success by 5.40 points. Training comparisons remain descriptive because variants are measured on their own sampled tasks.

8 Conclusions and Limitations

We present TrustMed-RL, an evidence-grounded framework for end-to-end agentic diagnosis. Adaptive GiGPO with evidence-based diagnostic rewards improves diagnostic performance and evidence acquisition, with strong generalization across evaluated OOD benchmarks. Together, our framework and benchmark enable systematic investigation of how agents coordinate the diagnostic pathway and ground their decisions in evidence acquired through clinical interaction. Our pipeline provides a foundation for iterative agent self-improvement through the generation and curation of diagnostic interaction trajectories. Future work can use these trajectories to target observed failures and strengthen evidence-grounded decision-making, advancing more reliable clinical agents.

Limitations. Image-integrity detection declines after RL, plausibly reflecting joint-corruption supervision that underemphasizes axis-specific discrimination. Policy-dependent exposure further complicates comparisons. Low accuracy on empirically unsolvable cases also suggests a persistent capability bottleneck in the 8B policy.

AI USE STATEMENT

We used generative AI tools to improve the clarity and fluency of the manuscript, generate synthetic data, and provide limited assistance with code development. The synthetic data generation procedure is fully described in the paper. Our implementation builds on the open-source frameworks verl-agent (`https://github.com/langfengQ/verl-agent`), verl (`https://github.com/verl-project/verl`), and AgentGym-RL (`https://github.com/woooodyy/AgentGym-RL`). Two authors debugged the code to ensure reproducibility. The authors take full responsibility for the final manuscript, code, data, and reported findings, including content produced with AI assistance.

ETHICS STATEMENT

This study uses publicly available data and open-source models to investigate evidence-grounded clinical diagnosis. We acknowledge potential biases inherited from these resources and the risk of incorrect or unsupported diagnostic outputs. The proposed framework is intended for research purposes and should not substitute for professional medical judgment. Any clinical application would require further validation, appropriate clinical oversight, and safeguards to protect patient privacy and safety.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide a fully reproducible implementation of TrustMed-RL and the data used in our experiments in an anonymized repository provided as supplementary material. Training and evaluation details are given in Appendices B and E. These resources support independent replication and verification of our experimental results.

REFERENCES

Reem Agbareia, Mahmud Omar, Shelly Soffer, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. Visual-textual integration in llms for medical diagnosis: A preliminary quantitative analysis. 2025.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Erin P Balogh, Bryan T Miller, and John R Ball. *Improving diagnosis in health care*. National Academies Press (US), 2015.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

Xavier P Burgos-Artizzu, David Coronado-Gutiérrez, Brenda Valenzuela-Alcaraz, Elisenda Bonet-Carne, Elisenda Eixarch, Fatima Crispi, and Eduard Gratacós. Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Scientific Reports*, 10(1):10200, 2020.

Haichao Chen, Zhengyun Zhao, Songchi Zhou, Shikai Hu, Jinyuan Wang, Ye Jin, Xianghong Jin, Yih Chung Tham, Xiaofei Wang, Weizhi Ma, et al. Rarearena: a comprehensive benchmark dataset unveiling the potential of large language models in rare disease diagnosis. *The Lancet Digital Health*, 8(2), 2026.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, et al. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159, 2025.

Jun Cheng, Wei Huang, Shuangliang Cao, Ru Yang, Wei Yang, Zhaoqiang Yun, Zhijian Wang, and Qianjin Feng. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PloS one*, 10(10):e0140381, 2015.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6): 1045–1057, 2013.

Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, pp. 2048–2056. PMLR, 2020.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 758–759, 2009.

David Dai, Peilin Chen, Chanakya Ekbote, and Paul Liang. Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training. *Advances in Neural Information Processing Systems*, 38:37406–37453, 2026.

Hongxin Ding, Baixiang Huang, Yue Fang, Weibin Liao, Xinke Jiang, Jinyang Zhang, Yinghao Zhu, Zheng Li, Liantao Ma, Junfeng Zhao, et al. Promed: Shapley information gain guided reinforcement learning for proactive medical llms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32481–32515, 2026.

Arthur S Elstein and Alan Schwarz. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *Bmj*, 324(7339):729–732, 2002.

Nastaran Enshaei, Anastasia Oikonomou, Moezedin Javad Rafiee, Parnian Afshar, Shahin Heidarian, Arash Mohammadi, Konstantinos N Plataniotis, and Farnoosh Naderkhani. Covid-rate: an automated framework for segmentation of covid-19 lesions from chest ct images. *Scientific Reports*, 12(1):3212, 2022.

Zhiting Fan, Xuxiang Ren, Yuan Wang, Ruizhe Chen, Zijie Meng, Keli Hu, Jiaxiang Liu, Jian Wu, Weiqi Zhai, Hongxia Xu, and Zuozhu Liu. MedEnv: Scaling Multimodal Virtual Medical Environments for Long-Horizon Diagnosis . In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, volume LNCS 16886. Springer Nature Switzerland, September 2026.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026a.

Yichun Feng, Jiawei Wang, Lu Zhou, Zhen Lei, and Yixue Li. Doctoragent-rl: A multi-agent collaborative reinforcement learning system for multi-turn clinical dialogue. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16952–16956. IEEE, 2026b.

Farieda Gaber, Maqsood Shaik, Fabio Allega, Agnes Julia Bilecz, Felix Busch, Kelsey Goon, Vedran Franke, and Altuna Akalin. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, 8(1):263, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Lecheng Gong, Weimin Fang, Ting Yang, Dongjie Tao, Chunxiao Guo, Peng Wei, Bo Xie, Jinqun Guan, Zixiao Chen, Fang Shi, et al. Meddialogrubrics: A comprehensive benchmark and evaluation framework for multi-turn medical consultations in large language models. *arXiv preprint arXiv:2601.03023*, 2026.

MIaMIA Group. PHE-SICH-CT-IDS. figshare, August 2023. URL https://figshare.com/articles/dataset/PHE-SICH-CT-IDS/23957937.

Yu Gu, Jingjing Fu, Xiaodong Liu, Jeya Maria Jose Valanarasu, Noel CF Codella, Reuben Tan, Qianchu Liu, Ying Jin, Sheng Zhang, Jinyu Wang, et al. The illusion of readiness in health ai. *arXiv preprint arXiv:2509.18234*, 2025.

Ibrahim Ethem Hamamci, Sezgin Er, Enis Simsar, Atif Emre Yuksel, Sadullah Gultekin, Serife Damla Ozdemir, Kaiyuan Yang, Hongwei Bran Li, Sarthak Pati, Bernd Stadlinger, et al. Dentex: An abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic x-rays. *arXiv preprint arXiv:2305.19112*, 2023.

Ronald M Harden. What is an osce? *Medical teacher*, 10(1):19–22, 1988.

Sihan Huang. FOCUS-7D_Public. figshare, February 2026. URL https://figshare.com/articles/dataset/FOCUS-7D_Public/31338496.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. Medvlthinker: Simple baselines for multimodal medical reasoning. *arXiv preprint*, 2025.

Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pp. 451–462. Springer, 2019.

Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. Prioritized level replay. In *International Conference on Machine Learning*, pp. 4940–4950. PMLR, 2021.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

Yung Hwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. Doctor-r1: Mastering clinical inquiry with experiential agentic reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 153920–153955, 2026.

Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

Rebecca Sawyer Lee, Jared A Dunnmon, Ann He, Siyi Tang, Christopher Re, and Daniel L Rubin. Comparison of segmentation-free and segmentation-dependent computer-aided diagnosis of breast masses on a public mammography dataset. *Journal of biomedical informatics*, 113:103656, 2021.

Pengyuan Li, Xiangying Jiang, Chandra Kambhamettu, and Hagit Shatkay. Compound image segmentation of published biomedical figures. *Bioinformatics*, 34(7):1192–1199, 2018.

Shuyue S Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024.

Li Lin, Meng Li, Yijin Huang, Pujin Cheng, Honghui Xia, Kai Wang, Jin Yuan, and Xiaoying Tang. The sustech-sysu dataset for automated exudate detection and diabetic retinopathy grading. *Scientific Data*, 7(1):409, 2020.

Amirreza Mahbod, Christine Polak, Katharina Feldmann, Rumsha Khan, Katharina Gelles, Georg Dorffner, Ramona Woitek, Sepideh Hatamikia, and Isabella Ellinger. Nuinsseg: A fully annotated dataset for nuclei instance segmentation in h&e-stained histological images. *Scientific Data*, 11 (1):295, 2024.

Sílvia Mamede, Tamara Van Gog, Kees Van Den Berge, Remy MJP Rikers, Jan LCM Van Saase, Coen Van Guldener, and Henk G Schmidt. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *Jama*, 304(11):1198–1203, 2010.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 12076–12100, 2023.

Coşku Öksüz. KidneyStoneTR: A Patient-Level Split Abdominal CT Dataset for Kidney Stone Detection. figshare, March 2026. URL `https://figshare.com/articles/dataset/ROI-Guided_Abdominal_CT_Dataset_for_Kidney_Stone_Detection/31156783`.

Maxim Popov, Akmaral Amanturdieva, Nuren Zhaksylyk, Alsabir Alkanov, Adilbek Saniyazbekov, Temirgali Aimyshev, Eldar Ismailov, Ablay Bulegenov, Arystan Kuzhukeyev, Aizhan Kulanbayeva, et al. Dataset for automatic region-based coronary artery disease diagnostics using x-ray angiography images. *Scientific data*, 11(1):20, 2024.

Khaled Saab, Chunjong Park, Tim Strother, Jan Freyberg, David GT Barrett, Yong Cheng, Wei-Hung Weng, David Stutz, Nenad Tomasev, Anil Palepu, et al. Advancing conversational diagnostic ai with multimodal reasoning. *Nature Medicine*, pp. 1–11, 2026.

Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pp. 379–387. Springer, 2017.

Akshat Sanghvi, Naren Akash, Raza Imam, Amit Sharma, and Mohit Jain. Medxagent: Multi-agent consultation for interactive medical diagnosis. *arXiv preprint arXiv:2606.03416*, 2026.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Md. Arham Tabib, Sumyia Sabrin Liza, Md. Hasan Imam Bijoy, Md. Zahid Hasan, and Md. Rafiqul Islam Khan. OrthoFrac-XR: A Clinically Validated Multimodal High-Resolution X-ray Imaging Dataset for Bone Fracture Detection and Localization. figshare, April 2026. URL https://figshare.com/articles/dataset/_b_OrthoFrac-XR_b_A_Clinically_Validated_Multimodal_High-Resolution_X-ray_Imaging_Dataset_for_Bone_Fracture_Detection_and_Localization/32021085.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, 2025.

Thomas LA van den Heuvel, Dagmar de Bruijn, Chris L de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2d ultrasound images. *PloS one*, 13(8): e0200412, 2018.

Jasper W van der Graaf, Miranda L van Hooff, Constantinus FM Buckens, Matthieu Rutten, Job LC van Susante, Robert Jan Kroeze, Marinus de Kleuver, Bram van Ginneken, and Nikolas Lessmann. Lumbar spine segmentation in mr images: a dataset and a public benchmark. *Scientific Data*, 11(1):264, 2024.

Marc Sebastian von der Stück, Roman Vuskov, Simon Westfechtel, Robert Siepmann, Christiane Kuhl, Daniel Truhn, and Sven Nebelung. Visual large language models in radiology: A systematic multimodel evaluation of diagnostic accuracy and hallucinations. *Life*, 16(1):66, 2026.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Jingxiang Chen, Mohammad Kachuee, Teja Gollapudi, Yiwei Liao, Nicolas Scheffer, et al. Truthrl: Incentivizing truthful llms via reinforcement learning. *arXiv preprint arXiv:2509.25760*, 2025.

Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Xin Guo, Dingwen Yang, Chenyang Liao, Wei He, et al. Agentgym: Evaluating and training large language model-based agents across diverse environments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27914–27961, 2025.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Jiaqi Liu, Honglin Guo, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, et al. Agentgym-rl: An open-source framework to train llm agents for long-horizon decision making via multi-turn rl. In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*, 2026.

Jiao Xu et al. Pulsemind: A multi-modal medical model for real-world clinical diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026a. arXiv:2601.07344.

Shicheng Xu, Xin Huang, Zihao Wei, Liang Pang, Huawei Shen, and Xueqi Cheng. Dxdirector: an agentic large language model driving the full-process clinical diagnosis. *Nature Communications*, 2026b.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026c.

Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Jie Sun, Xiang Li, Jingxian He, Wenyue Hua, et al. Simulated patient systems powered by large language model-based ai agents offer potential for transforming medical education. *Communications Medicine*, 6(1):27, 2025.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. In *Findings of the association for computational linguistics: EMNLP 2023*, pp. 10859–10885, 2023.

Huan Zhang, Qianglin Liu, Xiaolin Han, Lijuan Niu, and Weidong Sun. Tn5000: An ultrasound image dataset for thyroid nodule detection and classification. *Scientific data*, 12(1):1437, 2025.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pp. 4442–4457, 2025.

A Data and Environment Construction

Synthetic case generation. Our cases are derived from RareArena (Chen et al., 2026), a benchmark unveil the potential of LLMs in rare disease diagnosis. The benchmark contains rare-disease case reports from PubMed Central (PMC) with Orphanet-normalised diagnoses. We pair each case with the figures and tables of its source article from the PMC Open Access subset, verify that every image is retrievable, and label images as clinical content versus charts, tables, or diagrams; 40,356 articles with at least one clinical image remain (118,988 figures, 2,907 table images). Following prior work that grounds simulated patients and their evaluation in atomic facts (Min et al., 2023; Li et al., 2024; Yu et al., 2025; Gong et al., 2026; Ding et al., 2026), we use `gpt-5.6-terra` to convert each case report, together with its figures, tables, and captions, into an OSCE-style record (Harden, 1988). The record separates (i) a *patient actor*, the patient-reportable information in eleven history sections (encounter context, demographics, chief complaint, history of present illness, denied symptoms, past medical, medication, surgical, family, and social history, review of systems), each written as numbered atomic facts of at most 40 words with optional onset, duration, frequency, and severity attributes, from (ii) a *hidden profile*, the findings the doctor must uncover through the workup, organized into physical examination, laboratory, physiologic, imaging, and pathology results, each entry bound to the figure panel that supports it (panel identifier, modality, view, stain). The annotator may use only the case text, captions, and visible image content; it never states the diagnosis, and every finding obtained after treatment or at follow-up is quarantined in a separate post-baseline section so that the profile reflects the initial diagnostic state. Outputs are schema-validated (section membership, fact-identifier contiguity, explicit negation of denied symptoms, panel-order integrity); a rejected record is regenerated once with the validator's message as feedback. Of 15,000 attempted cases, 14,995 passed validation.

Capability-based labeling. Inspired by TruthRL Wei et al. (2025), we use staged diagnostic probes to estimate case difficulty and initialize the curriculum. MedGemma-27B, GPT-4o, and Claude Sonnet 5 serve as diagnostic judges. Five closed-book Qwen3-VL-8B responses classify cases as *easy* (4–5 successes), *medium* (2–3), or *hard* (0–1), with success requiring a mean ontology-calibrated judge score ≥ 0.85. Cases with no initial success undergo further probing with Qwen3-VL-8B with retrieval, Qwen3-VL-32B closed-book, and Qwen3-VL-32B with retrieval, with five responses per stage. An escalation response i passes if

$$\text{pass}_i = \mathbf{1}\left[\text{strict normalized match} \;\vee\; \frac{1}{3}\sum_{j=1}^{3}\max\{t_{ij}, w_{\text{ont},i}\} \geq 0.55\right],$$

where t_{ij} is judge j's diagnostic score and $w_{\text{ont},i}$ is the deterministic ontology-relation floor. Cases with complete evidence but no success across all probes are *empirically unsolved* (OOK) under the tested models and budgets. Probe stage and success counts define 13 ordered curriculum strata $r \in \{1, \ldots, 13\}$, refined into scores $c = (r - 1 + \delta)/13 \in [0, 1]$, where δ is one minus the mean diagnostic score at the assigned stage, or one minus the best escalation score for OOK cases. Higher scores indicate greater difficulty. These scores initialize mastery estimates, which are updated from training outcomes to adapt case sampling to the policy's evolving diagnostic capability.

A.1 Image Preparation and Review

Panel extraction and annotation. Many figures mined from PubMed Central are compound plates whose adjacent panels contain radiographs, microscopy or pathology images, plots, or other heterogeneous content. Prior work therefore treats separation into individual panels as a prerequisite for reliable downstream use (Li et al., 2018). Serving an entire compound figure could expose relevant and unrelated evidence simultaneously or assign a single modality, anatomy, or laterality label to heterogeneous pixels. We therefore crop figures to the semantic-panel level to align the displayed image with the ordered study and its integrity annotations.

Starting from 44,277 registered case figures, we annotated each semantic panel with its panel identity, source order, modality, anatomic site, and laterality, linked these annotations to pixel coordinates, and exported more than 24,000 panel-level crops. Each crop underwent manual review and segmentation, followed by checks that its figure identifier matched the source figure and its description matched the visual content. All segmented figures then underwent full-gallery human review, with rejection at either the panel or figure level. Rejected figures were re-segmented and re-reviewed

over multiple rounds. During both training and evaluation, victim images—the case's own evidence that the policy must verify—are restricted to human-verified single-panel assets. The frozen test split has complete panel coverage: every multi-panel test figure was manually separated and reviewed.

Integrity-failure labeling and probe construction. We augment clean cases with two synthetic data-integrity failures: **modality mismatch** and **anatomy mismatch**. Each mismatch episode is constructed through a seeded corruption that replaces the ordered image with a donor panel whose report-derived attributes disagree with the order on the target axis and agree on all other defined image-derived integrity axes. The injected axis is recorded as the ground-truth label, and the safe target is to identify the inconsistency and defer to a clinician. Both victim and donor panels are annotated for modality and anatomy and undergo human review to verify these annotations. Following counterfactual augmentation and procedural variation, clean and corrupted realizations are resampled across visits so that case identity alone cannot predict deferral (Kaushik et al., 2019; Cobbe et al., 2020; Geirhos et al., 2020). These probes evaluate whether the agent verifies evidence integrity before using it for diagnosis.

Donor-bank review. The training and validation donor pools contain 2,420 and 780 images, respectively, all reviewed under the same protocol. To fill gaps in modality–anatomy coverage, we generated 48 synthetic donors using `gemini-3.1-flash-image`. Each passed automated modality and body-region checks and individual human review before inclusion. For evaluation, we use a separate frozen bank of 580 distinct donors, comprising 332 PMC figure crops and 248 native 2D images from 21 openly licensed public datasets outside PMC. The PMC donors are PMCID-disjoint from all training datasets and asset pools, while the public-dataset donors are disjoint from these sources by construction. Each donor retains its source identifier and attribution.

The public datasets cover radiography Jaeger et al. (2014); Tabib et al. (2026); Hamamci et al. (2023); Lee et al. (2021), CT Group (2023); Enshaei et al. (2022); Öksüz (2026), MRI Cheng et al. (2015); van der Graaf et al. (2024); Bernard et al. (2018); Clark et al. (2013), ultrasound Al-Dhabyani et al. (2020); Zhang et al. (2025); Huang (2026); Leclerc et al. (2019); van den Heuvel et al. (2018); Burgos-Artizzu et al. (2020), coronary angiography Popov et al. (2024), gastrointestinal endoscopy Jha et al. (2019), fundus photography Lin et al. (2020), and H&E histology Mahbod et al. (2024).

GPT-5 assessed each evaluation donor in isolation. Donors whose body-site family it identified with confidence ≥ 0.95 were eligible for anatomy swaps; the remaining donors were used only for modality swaps.

A.2 Retrieval Corpus and Search Tool

For the retrieval protocol, we follow Search-R1 Jin et al. (2025). A standalone retrieval server answers the policy's search calls through a /retrieve endpoint. Each call returns the top-3 passages inside ¡information¿ tags, retrieved tokens receive no policy gradient, and a case allows at most three searches. Instead of Search-R1's 2018 Wikipedia corpus (21M passages, E5 retriever), we build a medical corpus following MedRAG Zhao et al. (2025). It has three collections. (i) PubMed: the 2026 MEDLINE baseline plus daily updates through July 30, 2026, with one title-plus-abstract passage per article and deleted or abstract-less records removed (29,166,067 passages). (ii) StatPearls: chapters split into one passage per paragraph, titled by their section path (358,500 passages). (iii) The 18 MedQA textbooks (125,847 passages). This gives 29.65M passages in total. Every passage is embedded with the MedCPT article encoder Jin et al. (2023); for PubMed, 84.5% of the vectors are reused from NCBI's precomputed release and the remaining 4.52M are encoded locally. Each collection has a BM25 index and a FAISS index. The FAISS index is exact for the two small collections. For PubMed it is an IVF65536-PQ64 index that re-scores the top 1,000 candidates exactly, reaching recall@10 = 0.959 against exhaustive search. At query time, BM25 and dense candidates are fused with reciprocal rank fusion Cormack et al. (2009). To prevent answer leakage, the article each case was drawn from is excluded at query time: all 14,995 case PMCIDs are mapped to PubMed IDs. Dense search runs on CPU and no neural reranker is used, the served index takes about 130 GB.

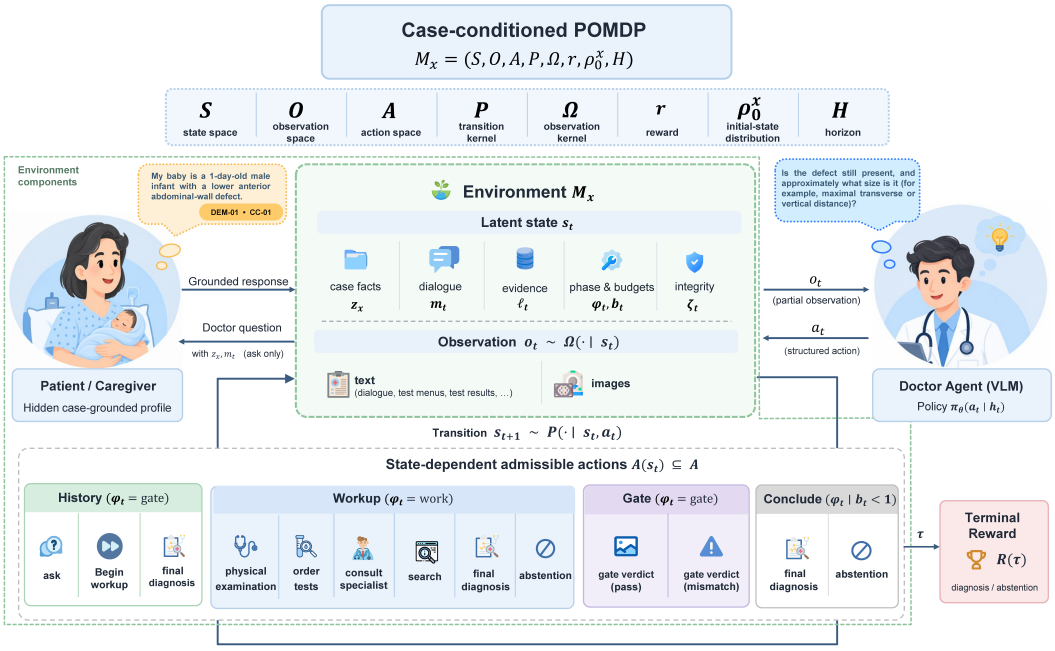


Figure 1: **Case-conditioned clinical POMDP.** The environment maintains the latent state s_t and releases a multimodal observation o_t to the physician. The physician selects an action a_t using its observable interaction context h_t, and the environment updates its state through P. The available actions depend on the current interaction phase φ_t, remaining budgets b_t, and evidence and integrity states. The patient simulator and clinical tools are environment components; only the physician policy π_θ is optimized. $R(\tau)$ denotes the terminal outcome reward.

B Training and Implementation Details

Case-Conditioned Clinical POMDP As illustrated in Figure 1, each clinical case defines a POMDP in which the environment maintains a latent state s_t and generates multimodal observations o_t, while the physician selects structured actions a_t based on its observable interaction history h_t. Admissible actions depend on the current interaction phase, remaining budgets, and evidence and integrity states, with only the physician policy π_θ optimized using the terminal outcome reward $R(\tau)$.

B.1 SFT Trajectory Generation

To initialize the policy for RL, we collect interactive demonstrations primarily from GPT-5-mini and Qwen3.8-Max, with GPT-4.1-mini simulating the standardized patient. Demonstrations span diagnostic difficulty and cover diagnosis, evidence-mismatch detection, and abstention under unresolved uncertainty. We apply trajectory-validity checks, outcome verification, and manual review. The diagnostic trajectories require an ontology-adjusted evaluation score of at least 0.85 to be accepted. The resulting dataset contains 1,287 trajectories: 860 diagnoses, 304 mismatch detections, and 123 uncertainty-based abstentions, which yields 24,329 state–action pairs, partitioned by case into approximately 95% training and 5% validation.

We fine-tune Qwen3-VL-8B-Instruct using LoRA ($r = 16$, $\alpha = 32$) with cross-entropy loss on teacher reasoning and action tokens only; patient responses and tool observations are excluded from the loss. Training uses a learning rate of 10^{-4}, global batch size 64, cosine decay with 3% warmup, and a maximum sequence length of 24,576 tokens for up to three epochs. The resulting model initializes the RL policy.

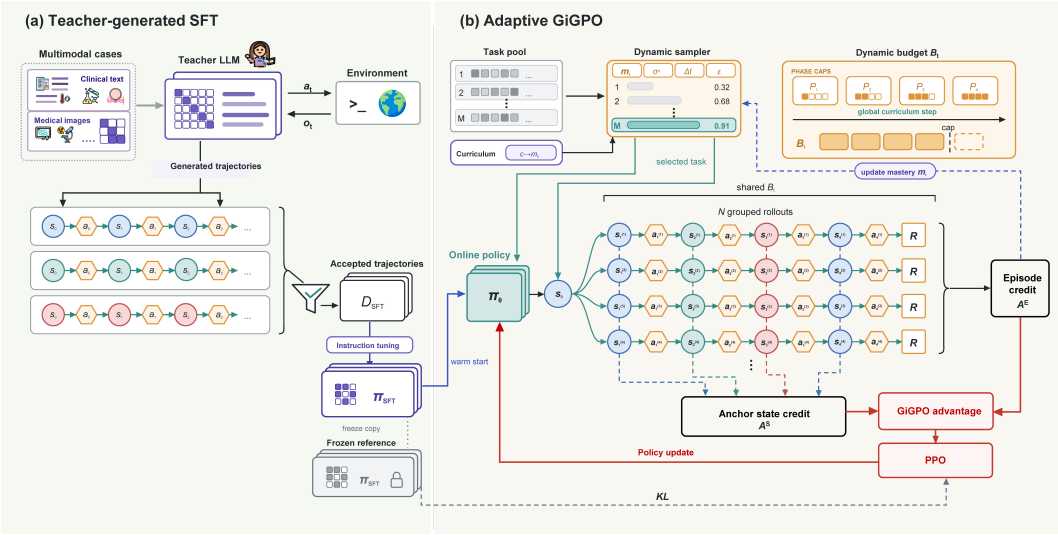


Figure 2: Training framework. **(a) Teacher-generated SFT.** Teacher models interact with the environment to generate multi-turn consultation trajectories. Quality-filtered trajectories form $\mathcal{D}_{\mathrm{SFT}}$; the resulting SFT policy initializes π_θ and serves as the frozen KL reference π_{ref}. **(b) Adaptive GiGPO.** A dynamic sampler combines outcome dispersion, staleness, and exploration to select instances admitted by the interaction curriculum. Each instance produces K trajectories under shared initial conditions and budgets. Episode-relative and step-relative advantages jointly guide policy optimization. Observed outcomes update sampling statistics and admission estimates, while a staged schedule increases the interaction-budget ceilings throughout the optimization steps.

B.2 Sampling and Interaction Curriculum

Phase schedule. The curriculum spans $N = 256$ training updates. At each update, we sample a batch of $B = 16$ training instances and collect a group of $K = 9$ trajectories per instance, yielding 144 trajectories. Sampling plans are refreshed every $C_{\mathrm{plan}} = 4$ updates using the completed sampling rounds from preceding blocks. The phase and interaction-budget ceilings are

$$\big(p(n), \mathbf{b}_{p(n)}\big) = \begin{cases} (1, (4, 10, 1)), & 0 \le n < 40, \\ (2, (6, 12, 2)), & 40 \le n < 88, \\ (3, (8, 14, 3)), & 88 \le n < 152, \\ (4, (10, 20, 3)), & 152 \le n < 256, \end{cases} \tag{8}$$

where $\mathbf{b}_p = (b_p^{\mathrm{ask}}, b_p^{\mathrm{work}}, b_p^{\mathrm{search}})$ specifies question, workup, and search ceilings. Phase transitions follow this fixed schedule; online mastery estimates do not advance them. The final ceiling equals the evaluation budget.

Instance admission and refinement. Let $(\hat{w}_x, \hat{s}_x)$ denote the offline estimates of required non-search workup actions and searches. We use the stored estimates for diagnosis and corrupted-evidence instances and $(\hat{w}_x, \hat{s}_x) = (2, 0)$ for out-of-knowledge instances. Since a search consumes both a workup turn and a search allowance, the initial admission class is

$$h_{x0} = \min\big\{p : \hat{w}_x + \hat{s}_x \le b_p^{\mathrm{work}},\ \hat{s}_x \le b_p^{\mathrm{search}}\big\}. \tag{9}$$

Instances fitting no phase are excluded. For trajectory (x, j, i), let w_{xji} be its observed workup usage, including searches, and $s_{xji}^{\mathrm{search}}$ its search count. Define its smallest usage-compatible phase as

$$p_{xji}^{*} = \min\big\{p : w_{xji} \le b_p^{\mathrm{work}},\ s_{xji}^{\mathrm{search}} \le b_p^{\mathrm{search}}\big\}.$$

Using the normalized sampler score $\widetilde{R}_{xji}$ from Eq. 4, we accumulate certificates from successful trajectories and update the admission class by

$$\begin{aligned} C_{xp,j} &= C_{xp,j-1} + \sum_{i=1}^{K} \mathbf{1}\{\widetilde{R}_{xji} \geq \zeta\}\mathbf{1}\{p^*_{xji} = p\}, \\ h_{xj} &= \min\big(\{h_{x,j-1}\} \cup \{p : C_{xp,j} \geq \kappa\}\big), \end{aligned} \tag{10}$$

where $C_{xp,0} = 0$, $\zeta = 0.5$, and $\kappa = 2$. Certificates accumulate at the exact smallest fitting phase, persist across sampling rounds, and may include multiple trajectories from one group. The resource test uses workup and search counts only. Updating h_{xj} leaves the offline requirements unchanged.

Shared budget randomization. For sampling round (x, j), let $p = p(n_{xj})$ and define

$$\begin{aligned} \underline{b}^{\text{ask}}_{xj} &= \max\{1, b^{\text{ask}}_p - \Delta_{\text{ask}}\}, \\ \underline{b}^{\text{work}}_{xj} &= \min\big\{b^{\text{work}}_p, \max\{2, b^{\text{work}}_p - \Delta_{\text{work}}, \hat{w}_x + \hat{s}_x\}\big\}, \end{aligned}$$

where $\Delta_{\text{ask}} = 1$ and $\Delta_{\text{work}} = 2$. The realized ceilings are sampled uniformly over inclusive integer intervals:

$$b^a_{xj} \sim \text{Unif}_{\mathbb{Z}}[\underline{b}^a_{xj}, b^a_p], \quad a \in \{\text{ask}, \text{work}\}, \qquad b^{\text{search}}_{xj} = b^{\text{search}}_p. \tag{11}$$

The same realized ceilings apply to all K trajectories in the group. Randomization remains active in the final phase.

Task-type and reward schedules. Let $g(n) = \min\{1, n/(0.3333N)\}$. The target task-type proportions and out-of-knowledge reference reward follow

$$\begin{aligned} \vartheta_{\text{ook}}(n) &= 0.15, \qquad \vartheta_{\text{corrupt}}(n) = 0.10 + 0.15g(n), \\ \vartheta_{\text{answer}}(n) &= 1 - \vartheta_{\text{ook}}(n) - \vartheta_{\text{corrupt}}(n), \\ r_{\text{ook}}(n) &= 0.15 + 0.30g(n). \end{aligned} \tag{12}$$

The sampler uses $R^{\text{ref}}_{\text{answer}}(n) = 1$, $R^{\text{ref}}_{\text{corrupt}}(n) = 0.9$, and $R^{\text{ref}}_{\text{ook}}(n) = r_{\text{ook}}(n)$ in Eq. 4. These are reference scales; normalized sampler scores are clipped to $[0, 1]$. Task-type proportions are evaluated at the start of each planning block, whereas r_{ook} is evaluated at the update of each sampling round.

Sampling and allocation. The dynamic sampler uses $\underline{m} = 0.05$, $\eta = 0.5$, $\beta_{\text{r}} = 0.3$, and $\rho = 0.2$ in the outcome-adaptive sampling rule of Section 5. For a draw requesting b instances from eligible set $\mathcal{E}$, let $b' = \min\{b, |\mathcal{E}|\}$. If $b' > 0$, the first

$$b_{\text{unif}} = \max\{1, \text{round}(\xi b')\}, \qquad \xi = 0.1,$$

selections are uniform; the remaining selections use the priority weights. Sampling is without replacement, with weights renormalized over the remaining candidates and ranks fixed within each draw.

Each planning block contains $C_{\text{plan}}B = 64$ instance selections. Let $\mathcal{T}_n$ be the retired, horizon-admitted instances at block initialization. We reserve

$$B_{\text{replay}} = \min\{\text{round}(0.05\, C_{\text{plan}}B), |\mathcal{T}_n|\}$$

slots for uniform replay probes. The remaining slots are apportioned according to $\vartheta_k(n)$ by integer floors and largest fractional remainders, after renormalizing over nonempty active task-type strata. Type slots and replay slots are shuffled and partitioned among the block's updates. Unfillable slots are reassigned to eligible active instances of any type when available.

Exposure and instance retention. Each training instance x is limited to $V_{\max} = 3$ sampling rounds, including replay probes and rounds already reserved in the current plan. Each update contains at most one instance of any clinical case. The retirement rule uses

$$(L_{\text{ret}}, \mu_{\text{ret}}, \sigma_{\text{ret}}, \mu_{\text{resume}}) = (2, 0.9, 0.05, 0.8).$$

A replay probe with $\mu_{xj} < \mu_{\text{resume}}$ clears retirement, but does not override the sampling-round limit. The additional low-performance shelving criterion requires three sampling rounds; under $V_{\max} = 3$, it does not further restrict future sampling. Within a planning block, the stored priority scores $u_{x,J_x(n)}$ and retirement states remain fixed while the sampling-round reservations used by $V_x(n)$ are updated.

B.3 Optimization and Computing Configurations

Hyperparameters. All RL runs are initialized from the SFT checkpoint and share a single configuration. The actor learning rate is set to 1×10^{-6} using AdamW with weight decay 0.01 and gradient-norm clipping at 1.0. The mini-batch size is 144, which corresponds to about 20 gradient updates per iteration within a single PPO epoch, and the KL-divergence loss coefficient is set to 0.01, using the low-variance k_3 estimator against the frozen SFT reference policy. We adopt asymmetric clipping with $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$, token-level loss averaging, and no entropy bonus. The rollout temperature is set to 1.0 (top-$p = 1.0$), while the validation temperature is set to 0.0 (greedy decoding); the maximum prompt and per-turn response lengths are $15{,}360$ and $1{,}024$ tokens, respectively. For GiGPO, both the episode-level advantage A^E and the step-level advantage A^S are normalized by their group mean and standard deviation, with the standard deviation floored at 0.125 to avoid degenerate groups. To scale up the absolute advantage of trajectories that terminate in a full-credit correct diagnosis, `DX_ADV_SCALE` is set to 1.2. The weighting coefficient ω of the step-level advantage A^S is fixed at 1 without further tuning, and the discount factor γ is set to 0.99.

Computing Details. All experiments are run on a single node with $4\times$ NVIDIA B300 GPUs. Training is implemented on top of `verl-agent`: the actor is trained with fully sharded data parallelism in bf16 with gradient checkpointing and padding removal, the frozen reference policy is kept in bf16 with parameter offloading, and rollouts are generated by a co-located vLLM engin. We use full-parameter updates with a micro-batch size of 3 sequences per GPU. Each run trains for 256 iterations; the policy is evaluated every 20 iterations on validation episodes, and checkpoints are saved at the same interval. One iteration takes 14.4 minutes on average, a full 256-iteration run takes approximately hours of 250 GPU-hours. Following The simulated patient is served by `gpt-4.1-mini` at temperature 0, and the reward judge uses the same model, with judge outputs cached per episode.

C Reward Specification

Diagnosis credit. During training, a single GPT-4.1-mini judge compares the final diagnosis with the reference using the semantic matching procedure in Section 7.1. The weights for FULL, CORE, and PARTIAL matches are 1.0, 0.85, and 0.50, respectively. Let m_i and n_i be the numbers of reference and predicted conditions, and let $\widetilde{s}_i$ be the matched mass after guarded ontology augmentation. The score is

$$u_i = \frac{\widetilde{s}_i}{\widetilde{s}_i + 0.30(n_i - \widetilde{s}_i) + 0.70(m_i - \widetilde{s}_i)} = \frac{\widetilde{s}_i}{0.30n_i + 0.70m_i}. \tag{13}$$

Thus, missed conditions receive greater weight than unsupported predictions in the Tversky denominator (Salehi et al., 2017). The ontology floor is consulted when the judge score is below 0.85. An accepted case-level ontology verdict raises existing match labels to at least CORE or FULL; if the judge found no match, one match at that level is inserted and the counts are recomputed. We threshold the resulting score to obtain diagnosis credit:

$$d_i = \begin{cases} 0, & u_i < 0.85, \\ 0.5, & u_i \geq 0.85 \text{ and the ontology strongly contradicts the match}, \\ 1, & \text{otherwise}. \end{cases} \tag{14}$$

The middle case applies only when both diagnoses are confidently resolved to unrelated disease entities without contrary ontology evidence. Failed judge parsing is retried once; if judging remains unavailable, a guarded deterministic ontology scorer assigns binary credit for an accepted same-disease match. This fallback is recorded separately.

Coverage adjustment. Let M_i be the number of image-bearing evidence records available in the case and L_i the number whose specialist reports have been released during the consultation. For $M_i > 0$, define

$$c_i = \min\left\{1, \frac{L_i}{M_i}\right\}, \qquad R_i^{\text{dx}} = d_i\,(0.10 + 0.90\sqrt{c_i})\,. \tag{15}$$

Coverage is measured over evidence records, not questions or tool calls; ordering or viewing an image alone does not count as a released report. Repeated requests do not increase the count of released

Table 3: Terminal rewards used in training. Diagnosis overrides are applied in the order described in the text. Invalid-action costs are separate.

Terminal outcome	R_i
Diagnosis, subject to the overrides below	R_i^{dx}
Incorrect diagnosis with no workup action	-0.05
Diagnosis after accepting a corrupted image batch	-0.10
Evidence Corruption detected with valid axis attribution	0.90
Evidence Corruption detected with invalid or missing axis attribution	0.65
False or unsupported evidence corruption alarm	-0.10
Deferral on an out-of-knowledge instance	$r_{\mathrm{ook}}(n)$
Out-of-knowledge deferral on any other task type	-0.05
Failure to produce a valid terminal action	-0.10

records. The adjustment is applied once, to the terminal diagnosis credit; intermediate evidence acquisition receives no separate positive reward. When no record-level expectation is available, the implementation uses the fraction of expected departments whose reports were released; if neither expectation is available, it sets $c_i = 1$. For $d_i = 1$, zero, one-quarter, and full coverage yield $R_i^{\mathrm{dx}} = 0.10, 0.55$, and 1.00, respectively.

Terminal outcomes. Table 3 specifies the terminal reward R_i. For corruption detection, the reported axes must form a nonempty subset of the true corruption axes to receive 0.90; detecting corruption without satisfying this condition receives 0.65. The axis vocabulary is patient identity, modality, and anatomy. Out-of-knowledge deferral is rewarded only on instances assigned to that task category.

For diagnosis outcomes, the zero-workup penalty takes precedence. Otherwise, accepting an exposed corrupted image batch and subsequently diagnosing incurs -0.10. If corruption was scheduled but never delivered, the diagnosis reward is capped at $h(n) = \min\{0.40, r_{\mathrm{ook}}(n)\}$, except when all expected image-bearing records have been covered ($M_i > 0$ and $c_i = 1$). The capped reward is $\min\{R_i^{\mathrm{dx}}, h(n)\}$. An incorrect diagnosis after workup receives zero unless an override applies. Workup usage counts consumed workup turns, including rejected actions. The experiments use image corruption.

Deferral schedule. For the $N = 256$ training updates, indexed by $n = 0, \ldots, N-1$, the out-of-knowledge reward follows

$$r_{\mathrm{ook}}(n) = 0.15 + 0.30 \min\left\{1, \frac{n}{0.3333N}\right\}, \tag{16}$$

rounded to four decimal places in the implementation. It increases from 0.15 to 0.45 over approximately the first third of training, as specified in Appendix B.2.

Invalid actions and credit assignment. Each format violation incurs a fixed cost $\lambda_{\mathrm{inv}} = 0.01$, equal to 1% of the maximum diagnosis reward. This cost does not depend on the episode's realized diagnosis credit. With $v_{it} = 0$ for a penalized format violation and $v_{it} = 1$ otherwise, the scores supplied to the episode- and step-relative advantage estimators are

$$q_{it}^E = R_i - 0.01(1 - v_{it}), \qquad q_{it}^S = 0.99^{T_i - t} R_i - 0.01(1 - v_{it}). \tag{17}$$

As in the episode-level and step-level advantages A^E and A^S entering Eq. 2, the penalty is assigned to the offending action itself and is not propagated to preceding actions. The episode score used for logging and sampling is $R_i - 0.01 \sum_t (1 - v_{it})$ before sampler normalization and clipping. The diagnosis advantage coefficient $\alpha_{\mathrm{dx}} = 1.2$ in Eq. 2 acts after advantage normalization; it does not alter the terminal reward specified here.

D Additional Experiment results

D.1 Out-of-Distribution Evaluation

We evaluate OOD generalization on three interactive diagnostic benchmarks. MTMedDialog consists of multi-turn doctor–patient consultations spanning multiple disease categories (Feng et al.,

Table 4: Diagnostic performance on 2,500 clean cases per evaluation run. All scores are percentages. SFT, GiGPO, and GRPO each use three training seeds, with three separately seeded evaluation passes per trained model (nine runs per method). Where reported, means and sample standard deviations (std) are computed across the three evaluation passes. $\Delta_{\mathrm{Acc}} = 100(\mathrm{Acc} - \mathrm{Acc@50})/\mathrm{Acc}$ and $\Delta_{\mathrm{strict}} = 100(\mathrm{Acc}^{\mathrm{strict}} - \mathrm{Strict@50})/\mathrm{Acc}^{\mathrm{strict}}$ measure relative accuracy reductions under the evidence-coverage requirement (lower is better). Non-diagnosis outcomes receive zero. Bold indicates the best result among open-weight models.

Model	Acc	$\mathrm{Acc}^{\mathrm{strict}}$	Prec.	Rec.	Acc@50	Strict@50	Δ_{Acc} (%)	Δ_{strict} (%)
Proprietary (closed-source) models								
GPT-5	54.2	22.0	42.9	50.0	43.2	17.7	20.3	19.5
GPT-4o	37.9	20.8	34.9	36.0	25.7	14.3	32.3	31.3
Open-weight models								
TrustMed-RL (GiGPO, 8B)	$\mathbf{37.1}_{\pm 1.1}$	$15.3_{\pm 1.0}$	$\mathbf{31.8}_{\pm 1.1}$	$\mathbf{34.6}_{\pm 1.1}$	$\mathbf{32.5}_{\pm 1.3}$	$\mathbf{13.5}_{\pm 0.6}$	12.3	11.9
TrustMed-RL (GRPO, 8B)	$34.5_{\pm 1.6}$	$13.6_{\pm 1.9}$	$28.9_{\pm 1.9}$	$32.1_{\pm 1.5}$	$30.8_{\pm 1.3}$	$12.2_{\pm 1.6}$	**10.7**	**10.9**
TrustMed-SFT (8B)	$24.7_{\pm 0.6}$	$9.0_{\pm 0.3}$	$19.0_{\pm 0.4}$	$22.9_{\pm 0.6}$	$18.8_{\pm 0.4}$	$6.9_{\pm 0.3}$	23.9	23.2
TrustMed-SFT (2B)	$13.0_{\pm 0.1}$	$3.8_{\pm 0.3}$	$9.3_{\pm 0.2}$	$12.3_{\pm 0.2}$	$8.9_{\pm 0.1}$	$2.7_{\pm 0.4}$	31.1	30.3
PulseMind-72B	$21.8_{\pm 0.3}$	$10.1_{\pm 0.3}$	$19.2_{\pm 0.2}$	$20.9_{\pm 0.1}$	$11.1_{\pm 0.1}$	$5.5_{\pm 0.3}$	49.4	45.8
HealthGPT-Pro-27B	$34.5_{\pm 0.7}$	$\mathbf{18.1}_{\pm 0.9}$	$30.8_{\pm 1.0}$	$32.6_{\pm 0.8}$	$23.5_{\pm 0.8}$	$12.7_{\pm 0.8}$	31.8	30.0
Qwen3-VL-32B-Instruct	30.4	14.9	26.9	28.6	20.9	10.2	31.2	31.4
MedGemma-27B-it (Sellergren et al., 2025)	30.4	14.9	27.6	28.7	20.9	10.7	31.1	28.4
Hulu-Med-32B (Jiang et al., 2025)	23.4	10.3	19.8	22.1	14.4	6.7	38.4	35.3
MedVLThinker-32B-RL	21.1	10.6	19.3	20.0	12.6	6.4	40.0	39.2
Lingshu-32B (Xu et al., 2026c)	21.5	10.0	19.5	20.6	13.9	7.0	35.5	29.3
QoQ-Med-VL-32B Dai et al. (2026)	13.4	4.8	11.5	13.1	8.6	3.7	36.3	21.8
Qwen3-VL-8B-Instruct (Bai et al., 2025)	$16.3_{\pm 0.3}$	$6.4_{\pm 0.5}$	$13.7_{\pm 0.2}$	$15.3_{\pm 0.3}$	$6.6_{\pm 0.2}$	$2.7_{\pm 0.1}$	59.2	58.0
HuatuoGPT-Vision-7B (Zhang et al., 2023)	$7.3_{\pm 0.3}$	$2.5_{\pm 0.2}$	$6.2_{\pm 0.4}$	$6.8_{\pm 0.3}$	$2.4_{\pm 0.1}$	$0.9_{\pm 0.1}$	67.2	63.2
MedVLThinker-7B Huang et al. (2025)	$6.8_{\pm 0.4}$	$2.3_{\pm 0.1}$	$6.2_{\pm 0.5}$	$6.5_{\pm 0.5}$	$3.4_{\pm 0.1}$	$1.3_{\pm 0.1}$	50.5	43.9
Qwen3-VL-2B	$2.0_{\pm 0.1}$	$0.5_{\pm 0.1}$	$1.8_{\pm 0.1}$	$1.9_{\pm 0.1}$	$0.6_{\pm 0.2}$	$0.3_{\pm 0.1}$	68.0	51.3

2026b). AgentClinic-MedQA transforms USMLE-style MedQA cases into sequential clinical encounters in which the agent gathers information before making a diagnosis, while AgentClinic-NEJM contains multimodal diagnostic cases adapted from *New England Journal of Medicine* image challenges (Schmidgall et al., 2024). Unlike their original multiple-choice or static formulations where applicable, our evaluation requires models to interact with the simulated clinical environment and provide a free-form final diagnosis. For MTMedDialog, we apply the DoctorAgent-RL final-diagnosis prompt to obtain the final prediction.

Table 6: Out-of-distribution diagnostic accuracy (Acc, %). Subscripted $\pm$ values denote standard deviations. Bold scores indicate the highest mean for each dataset.

Model	**MedQA**	**NEJM**	**MTMedDialog**
HealthGPT-Pro-27B	$\mathbf{73.0}_{\pm 1.2}$	$\mathbf{35.6}_{\pm 3.4}$	$39.5_{\pm 48.9}$
Qwen3-VL-32B-Instruct	$65.3_{\pm 0.7}$	$31.4_{\pm 1.1}$	$28.3_{\pm 45.0}$
MedGemma-27B-it	$70.4_{\pm 1.0}$	$34.0_{\pm 1.5}$	$33.3_{\pm 47.1}$
Hulu-Med-32B	$55.1_{\pm 49.9}$	$22.2_{\pm 41.8}$	$29.1_{\pm 45.5}$
MedVLMThinker-32B-RL	$65.3_{\pm 1.9}$	$29.7_{\pm 4.0}$	$36.3_{\pm 48.1}$
Lingshu-32B	$56.2_{\pm 1.6}$	$24.1_{\pm 3.2}$	$31.3_{\pm 46.4}$
QoQ-Med-VL-32B	$52.6_{\pm 5.8}$	$17.5_{\pm 4.9}$	$37.0_{\pm 48.3}$
Qwen3-VL-8B-Instruct	$53.6_{\pm 1.6}$	$21.4_{\pm 2.5}$	$40.7_{\pm 49.1}$
HuatuoGPT-Vision-7B	$36.5_{\pm 3.1}$	$14.2_{\pm 1.1}$	$20.4_{\pm 40.3}$
MedVLMThinker-7B	$36.0_{\pm 2.6}$	$13.5_{\pm 0.6}$	$32.9_{\pm 47.0}$
TrustMed-RL (GiGPO, 8B)	$62.2_{\pm 2.1}$	$26.2_{\pm 2.1}$	$48.7_{\pm 1.4}$
TrustMed-RL (GRPO, 8B)	$62.6_{\pm 2.9}$	$23.9_{\pm 0.8}$	$\mathbf{49.5}_{\pm 0.6}$

D.2 Ablation Results

Table 7 reports the held-out ablation results discussed in Section 7.3.

Table 5: Agentic behavior for models in Table 4. Order and Findings measure mean test-record ordering and findings coverage over 2,452 eligible clean cases. Findings includes laboratory, physiologic, imaging, and pathology reports, excluding raw images. Prod. counts actions yielding new history facts, images, or findings; Workup counts workup turns, both per clean episode. Search measures clean episodes with any search; Invalid measures invalid clean-episode turns. Modality and Anatomy Catch measure exposure-conditioned mismatch detection. False Alarm is the false-positive rate on matched clean-image gate-replay controls. All metrics except Prod. and Workup are percentages; — denotes unavailable results.

Model	Order	Findings	Prod.	Workup	Search	Invalid	Modality Catch	Anatomy Catch	False Alarm
Proprietary (closed-source) models									
GPT-5	72.4	62.8	8.5	12.8	63.2	0.0	97.8	88.7	4.8
GPT-4o	62.0	53.5	6.9	10.9	75.6	0.0	95.4	85.7	6.4
Open-weight models									
TrustMed-RL (GiGPO, 8B)	$83.9_{\pm 4.0}$	$73.2_{\pm 3.9}$	$8.5_{\pm 0.4}$	$17.8_{\pm 0.8}$	$93.0_{\pm 2.5}$	$4.0_{\pm 1.1}$	$82.2_{\pm 1.1}$	$52.6_{\pm 4.2}$	$2.2_{\pm 0.8}$
TrustMed-RL (GRPO, 8B)	$\mathbf{84.7}_{\pm 1.3}$	$\mathbf{73.9}_{\pm 1.1}$	$\mathbf{9.1}_{\pm 0.2}$	$18.3_{\pm 0.6}$	$89.3_{\pm 4.7}$	$2.6_{\pm 0.2}$	$85.4_{\pm 7.9}$	$60.6_{\pm 2.4}$	$2.4_{\pm 0.5}$
TrustMed-SFT (8B)	$66.4_{\pm 0.8}$	$57.1_{\pm 0.7}$	$7.7_{\pm 0.0}$	$10.2_{\pm 0.1}$	$53.2_{\pm 1.4}$	$\mathbf{0.6}_{\pm 0.0}$	$92.2_{\pm 1.9}$	$\mathbf{73.3}_{\pm 1.4}$	$3.9_{\pm 0.8}$
TrustMed-SFT (2B)	$59.4_{\pm 0.1}$	$51.3_{\pm 0.2}$	$7.2_{\pm 0.0}$	$11.1_{\pm 0.1}$	$50.7_{\pm 1.5}$	$6.9_{\pm 0.1}$	$80.8_{\pm 5.7}$	$36.2_{\pm 10.1}$	$5.0_{\pm 0.4}$
PulseMind-72B	$45.4_{\pm 0.6}$	$37.6_{\pm 0.6}$	$5.1_{\pm 0.0}$	$7.2_{\pm 0.0}$	$36.6_{\pm 1.7}$	$5.3_{\pm 0.0}$	—	—	—
HealthGPT-Pro-27B	$61.6_{\pm 0.3}$	$53.3_{\pm 0.3}$	$6.3_{\pm 0.0}$	$12.3_{\pm 0.7}$	$53.2_{\pm 4.5}$	$17.0_{\pm 3.8}$	$82.0_{\pm 3.4}$	$70.9_{\pm 0.7}$	$5.9_{\pm 0.6}$
Qwen3-VL-32B-Instruct	55.0	47.1	6.0	11.5	61.3	5.2	83.6	68.8	3.9
MedGemma-27B-it	61.2	52.9	6.7	18.1	77.8	41.0	42.9	61.1	7.2
Hulu-Med-32B	49.9	42.6	5.9	11.8	51.4	15.4	51.2	40.8	2.7
MedVLThinker-32B-RL	52.5	45.2	6.1	9.5	53.2	6.1	78.5	59.3	6.9
Lingshu-32B	52.5	45.2	6.0	10.3	43.2	7.4	48.6	56.2	4.7
QoQ-Med-VL-32B	53.6	46.8	5.8	12.6	80.4	10.3	36.4	37.3	3.1
Qwen3-VL-8B-Instruct	$36.1_{\pm 1.0}$	$30.8_{\pm 0.8}$	$4.7_{\pm 0.0}$	$5.6_{\pm 0.1}$	$18.1_{\pm 0.9}$	$7.1_{\pm 0.3}$	$\mathbf{92.9}_{\pm 5.0}$	$57.0_{\pm 3.6}$	$4.6_{\pm 2.3}$
HuatuoGPT-Vision-7B	$33.9_{\pm 0.3}$	$26.4_{\pm 0.2}$	$2.9_{\pm 0.0}$	$15.4_{\pm 0.0}$	$36.5_{\pm 0.5}$	$65.6_{\pm 0.1}$	$32.4_{\pm 6.2}$	$17.6_{\pm 3.2}$	$0.2_{\pm 0.4}$
MedVLThinker-7B	$44.7_{\pm 0.4}$	$35.9_{\pm 0.2}$	$4.6_{\pm 0.0}$	$16.4_{\pm 0.1}$	$74.2_{\pm 0.5}$	$37.2_{\pm 0.5}$	$76.9_{\pm 0.9}$	$44.0_{\pm 7.5}$	$10.9_{\pm 0.8}$
Qwen3-VL-2B	$32.3_{\pm 0.6}$	$26.7_{\pm 0.5}$	$3.3_{\pm 0.0}$	$17.0_{\pm 0.2}$	$6.8_{\pm 0.4}$	$63.7_{\pm 0.2}$	$53.1_{\pm 4.9}$	$41.4_{\pm 9.0}$	$1.1_{\pm 0.5}$

Table 7: Ablation study. Accuracy values are percentages; gains are percentage points relative to the named reference. ACC@3 and Strict@3 measure the proportion of cases correct in at least one of three evaluations under the corresponding criterion; each case contributes once. Base denotes Qwen3-VL Instruct. All RL variants use SFT initialization. Brackets show paired-case 95% intervals; the four SFT/RL gains also survive Bonferroni adjustment.

	Solvable				Unsolvable					
Model / variant	ACC	Strict	ACC@3	Strict@3	ACC	Strict	ACC@3	Strict@3	Reference	Δ solvable ACC [95% CI]
2B Base	2.27	0.60	6.21	1.65	0.44	0.09	1.07	0.27	-	-
2B + SFT	14.57	4.33	31.62	11.20	4.00	0.98	9.33	2.93	2B Base	+12.30 [11.28, 13.33]
8B Base	18.40	7.39	36.71	15.95	4.18	0.71	10.40	2.13	-	-
8B + SFT	27.89	10.40	48.94	21.65	7.64	2.04	17.33	5.07	8B Base	+9.49 [8.03, 10.92]
TrustMed-RL w/o step credit (GRPO)	39.98	17.77	60.00	32.28	11.02	3.91	21.87	8.53	8B SFT	+12.09 [10.59, 13.62]
TrustMed-RL (GiGPO)	**41.73**	**18.20**	**61.88**	**32.80**	10.22	3.20	20.53	7.47	8B SFT	**+13.84 [12.28, 15.39]**
8B w/o std. normalization	39.83	16.03	60.28	29.79	11.56	3.11	22.40	6.67	Full	-1.90 [-3.42, -0.41]
8B w/o priority sampling	39.95	15.51	59.58	29.74	9.78	2.84	18.13	7.47	Full	-1.77 [-3.29, -0.27]
8B w/o adaptive sampling	40.88	16.38	61.55	30.35	10.40	3.02	19.20	6.67	Full	-0.85 [-2.37, 0.66]
8B w/o adaptive sampling, horizon	39.98	14.43	61.22	27.53	11.29	2.40	22.40	5.33	Full	-1.74 [-3.25, -0.25]

We analyze training behavior over the final 40 updates (217–256). Table 8 compares task outcomes, integrity checks, and interaction patterns across ablations. Training task categories are distinct from evaluation difficulty strata: cases in the empirically unsolvable evaluation subset are still assessed for diagnostic performance.

Ablation analysis. Across the final 40 updates, the full method achieves the highest answer-task diagnostic success (28.01%) and hard-band success (26.46%). Removing sampling priority or all outcome-adaptive sampling lowers diagnostic success to 20.83% and 18.53%, respectively, and corruption-task success from 75.99% to 69.59% and 66.39%, despite comparable or greater aggregate evidence coverage. Jointly removing adaptive sampling and the horizon curriculum yields 21.84% diagnostic success and 62.19% corruption-task success, with longer responses (401.5 versus 368.9 tokens). Separately, over updates 1–40, retaining progressive horizons with adaptive sampling disabled reduces interaction length by 45.5% relative to full-budget training. Removing standard-deviation normalization or step-relative credit yields 19.90% and 22.61% diagnostic success. Although GRPO obtains a similar terminal reward (0.318 versus 0.317), the full method achieves higher diagnostic success, distinguishing diagnostic improvement from aggregate reward optimiza-

Table 8: Late-training ablations over updates 217–256. Values are equally weighted means of available per-update statistics on each variant's sampled training tasks, not held-out evaluation results. Percentage metrics use a 0–100 scale. Bold marks the highest outcome or lowest error mean; coverage and interaction/output lengths are descriptive.

Metric	Full	w/o priority sampling	w/o adaptive sampling	w/o adaptive sampling + horizon	w/o std	w/o A^S (GRPO)
Task outcomes						
Overall task success (%)	38.63	33.45	29.64	32.40	34.39	**38.72**
Diagnostic success: answer (%)	**28.01**	20.83	18.53	21.84	19.90	22.61
Hard-band diagnostic success (%)	**26.46**	15.16	16.30	18.27	17.47	21.92
Groups with ≥ 1 correct diagnosis (%)	**68.12**	58.37	60.25	67.23	61.57	60.98
Terminal reward	0.3174	0.2661	0.2391	0.2526	0.2797	**0.3180**
Integrity check						
Corruption-task success (%)	75.99	69.59	66.39	62.19	75.66	**80.22**
Exposed-mismatch detection (%)	67.98	60.96	59.77	55.05	69.57	**74.56**
False mismatch rate (%)	4.68	5.84	5.70	6.34	6.15	**4.09**
Interaction and output						
Evidence coverage (%)	66.32	68.26	66.51	69.50	61.85	66.26
Questions per episode	4.41	8.42	2.59	4.28	7.20	7.63
Workup turns per episode	13.86	14.87	14.48	16.04	14.11	14.87
Searches per episode	1.65	2.17	1.58	1.96	1.36	1.53
Total turns per episode	21.81	26.79	20.61	23.80	24.65	25.93
Output length (tokens/response)	368.9	342.9	386.4	401.5	341.9	377.6
Invalid actions per 100 turns	8.65	6.62	6.52	6.69	**4.18**	4.48

Configurations. *w/o priority sampling* removes selection priority while retaining other adaptive rules. *w/o adaptive sampling* disables all outcome-adaptive sampling while retaining staged horizons. *w/o adaptive sampling + horizon* additionally replaces staged budgets and admission with full-budget training from the start and removes budget jitter. *w/o std* uses mean-only normalization ($F_{\text{norm}} = 1$) for both advantage components. *w/o* A^S removes step-relative credit, yielding GRPO. Task-type and OOK-reward schedules are retained in the sampling and horizon ablations.

Metric scope. Diagnostic success uses all answer-task episodes as the denominator; the hard-band row follows the logged difficulty split. Group success is the observed proportion of diagnostic groups with at least one correct rollout among nine. Corruption-task success is the logged task-success rate over all corruption-task episodes, whereas exposed-mismatch detection is conditional on exposure; these metrics use different success and conditioning criteria. Interaction and coverage statistics aggregate all task kinds. Output length is the mean generated token count per doctor-policy call, including reasoning and action, not total tokens per episode. Undefined conditional rates are omitted rather than treated as zero. Invalid actions per 100 turns is 100 times the ratio of the window means of invalid actions and total turns.

tion. These observations support distinct roles for adaptive task allocation, progressive interaction budgets, and normalized two-level credit, with diagnostic gains not explained by aggregate evidence coverage alone.

E Evaluation Details

E.1 Diagnosis Matching and Judge Validation

Condition extraction. We decompose ground-truth and predicted diagnoses into condition sets according to clinical meaning. Subtype, site, etiology, pathology, grade, stage, severity, and complications are treated as qualifiers of the underlying disease unless independently asserted as separate diagnoses.

Semantic matching. We assess candidate prediction–reference pairs using the four-level ordinal rubric in Table 9, following partial-credit approaches to evaluating medical answers and diagnostic outcomes (Chen et al., 2025; Gaber et al., 2025; Lai et al., 2026; Mamede et al., 2010). We evaluated eight diagnosis judges under this rubric, with three separate calls per prediction–reference pair in each evaluation setting.

Table 9: Semantic diagnosis-matching rubric used by the LLM judges.

Level	Weight	Definition
FULL	$w_{\mathrm{F}} = 1.00$	Identifies the same fully specified disease. Synonyms, abbreviations, spelling variants, and clinically equivalent terminology are accepted.
CORE	$w_{\mathrm{C}} = 0.85$	Identifies the same underlying disease but omits or misstates one or more clinically relevant qualifiers, such as subtype, site, etiology, pathology, severity, grade, stage, or complication.
PARTIAL	$w_{\mathrm{P}} = 0.50$	Correctly identifies an explicit constituent of the reference diagnosis, such as its etiologic agent, disease-defining manifestation, or complication, without identifying the core disease.
NULL	$w_{\varnothing} = 0$	Has no valid semantic correspondence with the reference. Shared anatomy, nonspecific symptoms, broad disease-family membership, or general clinical similarity alone do not constitute a match.

Table 10: Judge and ontology-floor ablations. Entries report without $\rightarrow$ with the ontology floor. ORDO measures conformance to prespecified ontology-relation rewards; physician-inclusive evaluation measures agreement with a mixed human–model reference. Agreement denotes exact four-level label agreement. All settings use three passes.

	ORDO ($N = 100$)	Physician-inclusive ($N = 139$)	
Judge	MAE ↓	Agree. (%) ↑	MAE ↓
GPT-4.1-mini	0.169 → **0.017**	46.0 → 46.0	0.334 → 0.321
GPT-5-mini	0.098 → 0.020	51.1 → 51.1	0.269 → 0.262
GPT-4o	0.159 → 0.019	46.0 → 46.8	0.333 → 0.314
GPT-4o-mini	0.392 → **0.017**	42.4 → 43.9	0.377 → 0.351
Claude Sonnet 5	0.067 → 0.037	**55.4** → **53.2**	**0.249** → **0.257**
MedGemma-27B	0.118 → 0.025	51.8 → 50.4	0.264 → 0.265
Gemma-4-26B-A4B	**0.057** → 0.027	46.0 → 44.6	0.309 → 0.311
Baichuan-M2-32B	0.123 → 0.024	48.9 → 48.2	0.319 → 0.313

Table 11: Construction of the ORDO-derived judge-conformance set. Relation classes map deterministically to expected rewards. Narrower and broader pairs are opposite orientations of the same 20 hierarchy edges and are resampled jointly, yielding 80 bootstrap clusters. The fixed class mixture has mean target reward 0.625 and does not represent deployment prevalence.

Relation class	n	Prediction → reference construction	Target	Resampling unit
Synonym	20	Distinct ORDO synonym → preferred name	Full (1.00)	Pair
Narrower	20	Subtype → parent disorder	Core (0.85)	Hierarchy edge
Broader	20	Parent disorder → subtype	Core (0.85)	Hierarchy edge
Sibling subtype	10	Two non-ancestral subtypes of one disorder	Core (0.85)	Pair
Sibling group	10	Two disorders in one small ORDO group	None (0)	Pair
Unrelated	20	No shared clinical ancestor at any depth	None (0)	Pair

Offline ontology check. We supplement judge scores with a deterministic check against fixed local snapshots of MedGen, MONDO, and ORDO 4.9, enabling reproducible scoring without live queries. Diagnosis terms are resolved using names, synonyms, and ontology relations. Equivalence supported by at least two ontologies sets a full-credit floor ($r_{\mathrm{onto}} = 1.00$); single-ontology equivalence, subtype–disorder relations, and ORDO-typed sibling subtypes set a core-credit floor ($r_{\mathrm{onto}} = 0.85$). Directed hierarchy evidence takes precedence over equivalence. When full-term matching fails, a second check of the core disease terms can establish a core-credit floor. Other relations impose no floor. The adjusted reward is $r^{\star} = \max(\hat{r}, r_{\mathrm{onto}})$, where $\hat{r}$ is the judge reward and $r_{\mathrm{onto}} = 0$ when no floor applies.

Ontology conformance. We constructed 100 ORDO-derived prediction–reference pairs spanning synonyms, both directions of subtype–disorder relations, sibling subtypes, sibling disorders, and unrelated disorders (Table 11). The floor reduced mean absolute reward error (MAE) for all eight

Table 12: Class-wise ontology-floor ablation on 100 ORDO pairs. All judges use the final strict-entity rubric and are evaluated in three passes. Each $a \rightarrow b$ entry denotes without $\rightarrow$ with the ontology floor. Class columns report mean rewards, with targets in parentheses; MAE is the mean absolute reward error over all 100 pairs. Bold marks only the lowest overall MAE, including ties. E1 uses MedGen, MONDO, and ORDO; E2 removes ORDO. Because ORDO supplies both the pairs and part of the floor, this is partly a self-consistency test.

Judge	Synonym (1.00)	Narrower (0.85)	Broader (0.85)	Sibling subtype (0.85)	Sibling group (0.00)	Unrelated (0.00)
GPT-4.1-mini	0.878 → 1.000	0.680 → 0.850	0.680 → 0.850	0.085 → 0.850	0.000 → 0.170	0.000 → 0.000
GPT-5-mini	0.993 → 1.000	0.772 → 0.858	0.772 → 0.858	0.340 → 0.850	0.085 → 0.170	0.000 → 0.000
GPT-4o	0.935 → 1.000	0.730 → 0.858	0.722 → 0.850	0.000 → 0.850	0.085 → 0.170	0.000 → 0.000
GPT-4o-mini	0.485 → 1.000	0.340 → 0.850	0.340 → 0.850	0.000 → 0.850	0.000 → 0.170	0.000 → 0.000
Claude Sonnet 5	0.993 → 1.000	0.830 → 0.872	0.815 → 0.858	0.645 → 0.850	0.220 → 0.305	0.000 → 0.000
MedGemma-27B	0.887 → 1.000	0.762 → 0.858	0.797 → 0.858	0.370 → 0.850	0.135 → 0.220	0.000 → 0.000
Gemma-4-26B-A4B	0.978 → 1.000	0.823 → 0.865	0.772 → 0.858	0.765 → 0.850	0.135 → 0.220	0.000 → 0.000
Baichuan-M2-32B	0.970 → 1.000	0.722 → 0.850	0.730 → 0.858	0.255 → 0.850	0.050 → 0.220	0.000 → 0.000

Table 13: Native-rule fidelity of our re-implementations of published judge protocols and descriptive mapping to our reward scale. Native fidelity compares each implementation with the labels prescribed by its own protocol; class counts use modal labels and the parenthesized overall ranges span the three repeats. Cross-rubric reward discrepancy maps native outputs to our rewards and is not a measure of protocol accuracy.

Panel A: Native-rule fidelity by relation class.

Protocol	Synonym	Narrower	Broader	Sibling subtype	Sibling group	Unrelated	Overall	3/3 agreement
RareArena (2/1/0)	19/20	5/20	20/20	8/10	7/10	20/20	79/100 (78–80)	98/100
MeDxAgent (true/false)	19/20	13/20	16/20	10/10	10/10	20/20	88/100 (87–90)	93/100

RareArena prescribes label 2 for synonyms, label 1 for credited hypernyms, and 0 otherwise; MeDxAgent prescribes true for synonyms and more-specific subtype predictions and false otherwise. Structural MAE is the policy mismatch of a perfectly compliant implementation; observed MAE additionally reflects implementation deviations.

judges, narrowing the observed MAE range from 0.057–0.392 to 0.017–0.037 (Table 10). GPT-4.1-mini tied for the lowest MAE (0.017), while GPT-5-mini achieved 0.020. Across judges, synonym and sibling-subtype rewards reached their prescribed targets, and unrelated pairs retained zero reward; residual error was concentrated in sibling-group pairs (Table 12). These results support more consistent implementation of the reward rubric across judge backbones on this ontology-derived set.

Physician-inclusive evaluation. Two senior physicians annotated 200 pairs selected for the highest inter-judge disagreement. Together with annotations from `gpt-5.6-sol` and BaichuanM3-PLUS, we retained 139 pairs with a unique label supported by at least two of the four raters, including at least one physician. On this reference set, the floor reduced MAE for GPT-4.1-mini from 0.334 to 0.321 and for GPT-5-mini from 0.269 to 0.262, while preserving exact agreement at 46.0% and 51.1%, respectively (Table 10). GPT-5-mini had the second-lowest MAE among floored judges, close to Claude Sonnet 5 (0.257). Improvements on this disagreement-enriched, mixed human–model reference set were smaller and model-dependent. Together with their low ontology-conformance errors, these results support our use of GPT-4.1-mini and GPT-5-mini with the ontology floor.

Published judge protocols. We separately evaluated reimplementations of RareArena and MeDxAgent (Chen et al., 2026; Sanghvi et al., 2026) against labels prescribed by their own decision rules (Table 13). Native-rule fidelity was $79/100$ and $88/100$, respectively. These results assess implementation fidelity to each protocol, not diagnostic accuracy or out-of-distribution generalization.

F PROMPT TEMPLATES

We provide the prompt templates for the doctor agent, simulated patient, and diagnosis evaluator. The doctor receives prompts specific to history-taking, diagnostic workup, evidence-integrity verification, and final decision-making. Each doctor prompt requests a `<reasoning>` block followed by exactly one admissible `<action>`.

Doctor Prompt: History-Taking

You are an expert doctor interviewing a {patient|caregiver} in a clinical consultation.
Patient:
{demographics}
Questions used:
{step_count}/{max_asks}.

INTERVIEW RULES
• Ask ONE focused bundle of 2–4 tightly related items on one topic, in at most two sentences and a single question.
• Cover roughly in order: chief complaint; history of present illness (onset, duration, progression, character, associated symptoms, key negatives, aggravating/relieving factors, prior evaluation/treatment); past medical history; medications/allergies; surgery; family history; social/exposure history.
• "I don't know" = UNKNOWN, not negative; do not re-ask that topic. Never repeat known or semantically equivalent questions. Stop when required domains are covered or unanswerable.

Known facts:
{known_facts}
Unanswerable topics:
{unknown_topics}
Consecutive answers with no new information:
{no_new_fact_streak}
Conversation:
{conversation}

Your admissible actions are:
<action>ask: [one focused question bundle]</action>
<action>final_diagnosis: [the most likely diagnosis]</action>
<action>begin_workup</action>

Use final_diagnosis only when current information is sufficient without further history, examination, testing, or search. begin_workup permanently ends the interview and opens the workup phase.

<reasoning>Identify the highest-priority missing information and what would change your differential, avoiding known or unanswerable topics.</reasoning>
<action>Exactly one admissible action.</action>

Figure 3: Doctor prompt for history-taking.

Doctor Prompt: Diagnostic Workup

You are an expert doctor in the diagnostic workup phase. History-taking is complete.

GENERAL RULE
Choose the action with the highest expected diagnostic value: support or refute a leading diagnosis, assess severity or complications, or rule out a must-not-miss condition. Work systematically; under-ordering can miss the diagnosis.

Your admissible actions are:
<action>physical_examination: 1-{max_items} categories</action> — exact names from {pe_menu}
<action>order_tests: 1-{max_items} L1/L2 pairs</action> — exact pairs from {test_menu}
<action>consult_specialist: one specialist</action> — from {dept_menu}
<action>search: one clinical question in plain English</action>
<action>abstention: mismatch — [modality / anatomy]</action>
<action>abstention: out_of_knowledge</action>
<action>final_diagnosis: [the most likely diagnosis]</action>

Patient history:
{history_facts}
Workup actions/results:
{workup_log}
Already used:
{used_line}
Studies awaiting interpretation:
{awaiting_line}
Workup actions remaining:
{remaining_workup_turns}

<reasoning>Reason step-by-step about the leading diagnosis, key alternatives, and the single action that best advances or ends the workup.</reasoning>
<action>Exactly one admissible action.</action>

Figure 4: Doctor prompt for diagnostic workup.

Doctor Prompt: Evidence Integrity Gate

EVIDENCE INTEGRITY CHECK

Your patient:
{patient_tag} ({demographics})
You ordered:
{ordered}
Returned:
{asset_list}

Step 1 —For EACH returned image, state in your reasoning its modality and anatomical site, as read from the image alone.

Step 2 — Compare each image with the ordered category and the documented history. A mismatch requires a positive conflict: a wrong modality or anatomy.

Reply with one
<reasoning> block followed by exactly one action:
<action>gate_verdict: pass</action>
<action>gate_verdict: mismatch — assets: ASSET_IDS; axes: modality|anatomy</action>

Figure 5: Doctor prompt for evidence-integrity verification.

Doctor Prompt: Conclusion

You are an expert doctor concluding a clinical consultation. History-taking and the diagnostic workup are complete; the workup budget is exhausted. No further workup actions remain.

Your admissible actions are:
<action>abstention: mismatch — [modality / anatomy]</action>
<action>abstention: out_of_knowledge</action>
<action>final_diagnosis: [the most likely diagnosis]</action>

Use mismatch only when a returned study or result is untrustworthy because it conflicts with the ordered modality/anatomy.

Use out_of_knowledge only when the evidence is trustworthy but the available history, workup, consultations, and search are insufficient for a safe diagnosis.

Patient history:
{history_facts}
Workup actions and returned results:
{workup_log}
Studies awaiting interpretation:
{awaiting_line}

<reasoning>Reason step-by-step about the leading diagnosis and key alternatives, weighed against the evidence gathered.</reasoning>
<action>Exactly one admissible terminal action.</action>

Figure 6: Doctor prompt for Conclusion

Patient Prompt Template

You are a standardized patient in a medical history-taking simulation.

You receive only: encounter_facts; patient_actor_facts; conversation_history; the Doctor's current question; and speaker_role (patient or caregiver). Each fact contains section, fact_id, atomic_fact, and optional onset, duration, frequency, or severity. Use only these supplied facts.

1. SPEAK NATURALLY. Use first person as the patient, or speak naturally about the patient from the caregiver's perspective. Usually answer in 1–4 concise sentences. Never mention fact IDs in the natural answer.

2. ANSWER THE COMPLETE QUESTION. The Doctor may ask 2–4 related items; answer every component supported by the supplied facts. For broad but clinically coherent questions, provide all directly relevant facts and do not unnecessarily withhold them.

3. STRICT FACT GROUNDING. Every clinical statement must be directly supported by a supplied fact. Preserve age/sex, location/laterality, timing/duration, severity/frequency, numerical values, family relationships, medication/procedure names, and positive/negative meaning. Do not invent symptoms, infer causes, guess a diagnosis, add unstated details, or combine facts into a new causal conclusion.

4. UNKNOWN IS NOT NEGATIVE. Say a finding is absent only when a fact explicitly states so. Otherwise say "I don't know," "I'm not sure," or "I don't remember being told that." For partially answerable questions, answer supported parts and list unsupported components in unknown_topics.

5. CONSISTENCY. Repeated or rephrased questions receive the same factual answer; reuse the same fact IDs when appropriate.

6. OPENING STATEMENT. With empty conversation_history and no Doctor question, give 1–2 sentences revealing at most two atomic facts: Encounter_Context first, then one central Chief_Complaint fact; otherwise fall back to up to two Chief_Complaint facts, then one History_of_Present_Illness fact. Reveal no additional demographics, past history, medications, or other details until asked.

7. FACT-ID OUTPUT. used_fact_ids lists every fact used in the current answer: existing IDs only, each once, in order of first appearance. Include previously revealed IDs again when they support the current answer; exclude facts considered but not stated.

Return JSON only:

```
{
  "answer": "Natural patient-facing response.",
  "used_fact_ids": ["CC-01", "HPI-01"],
  "unknown_topics": []
}
```

Figure 7: Simulated-patient prompt for fact-grounded responses.

Judge Prompt: Diagnosis Evaluation

You are a medical expert evaluating a predicted diagnosis against a ground-truth diagnosis.

Ground truth: {ground_truth}
Prediction: {predicted}

1. IDENTIFY CONDITIONS

Identify distinct medical conditions in each input by clinical meaning, not wording. Subtype, site, etiology, pathology, grade, stage, severity, and complication are qualifiers, not separate conditions unless they are independent diagnoses.

2. MATCH CONDITIONS

Use the best overall one-to-one assignment. Each condition can be matched at most once.

FULL: Same fully specified disease, including synonyms, abbreviations, spelling variants, and clinically equivalent wording.

CORE: Core disease identified, but clinically relevant qualifiers are missing or incorrect.

PARTIAL: Core disease not fully identified, but a meaningful diagnostic component is correctly identified, such as an organism, complication, manifestation, or syndrome component.

3. OUTPUT

Return only valid JSON:

```
{
  "gt_count": int,
  "pred_count": int,
  "matches": [
    {"ground_truth": "", "predicted": "", "level": ""}
  ]
}
```

Report all distinct condition counts. Each level must be "full", "core", or "partial". Include only valid matched pairs; omit unmatched conditions from "matches".

4. STRICT ENTITY RULE

Match only the SAME disease entity. Synonyms and site, form, or qualifier variants may match. A specific disease may match a generic ground-truth category by disease classification, not merely location.

Different diseases are NOT a match at any level merely because they share anatomy, an organ system, histology, imaging features, or a differential diagnosis.

Figure 8: Evaluator prompt.

Table 14: Physician assessment of 200 automatically accepted diagnostic trajectories. Scores range from 1 to 5. We report the mean and sample standard deviation across trajectories, together with counts and percentages in the indicated score ranges.

Criterion	Mean $\pm$ SD	Scores 4–5, n (%)	Scores 1–2, n (%)
C1: Findings interpretation	4.33 ± 0.91	158 (79.0)	8 (4.0)
C2: Evidence use and revision	4.06 ± 0.93	144 (72.0)	11 (5.5)
C3: Evidential grounding	4.39 ± 0.92	166 (83.0)	12 (6.0)

G Reasoning Behavior

To assess if correct diagnoses were supported by grounded reasoning, we sampled 200 trajectories whose final diagnoses had been judged correct or largely correct by the automated evaluator. Two senior physicians reviewed a set of 200 trajectories, including the evidence obtained during the encounter, the turn-by-turn reasoning, the rationale in the final turn, and reference diagnoses. Each trajectory was rated on three dimensions: (1) **clinical findings interpretation**, assessing whether the findings used in reasoning were correctly interpreted; (2) **evidence use and diagnostic revision**, assessing whether the doctor agent formed a reasonable differential diagnosis and appropriately incorporated new or contradictory evidence; and (3) **evidential grounding**, assessing whether findings presented as observations were supported by information actually received.

Physician review results. All 200 trajectories received complete physician ratings. Table 14 summarizes the dimension-level results. Overall, 134 trajectories (67.0%) received scores of at least 4 on all three dimensions, whereas only 18 (9.0%) received a score of 1 or 2 on any dimension. Evidential grounding was strongest, with 83.0% of trajectories receiving scores of 4–5, followed by findings interpretation (79.0%) and evidence use and diagnostic revision (72.0%). These results indicate that, among diagnostically accepted trajectories, the model's rationales were generally well grounded in the observed evidence and clinically coherent, while evidence integration and diagnostic revision remained the primary source of residual reasoning errors.

Environment (Step 1)

You are an expert doctor interviewing a patient in a clinical consultation.
Patient demographics: 27-year-old male (PT-17244).
Patient: I've noticed a palpable mass in my right testicle for the past three months.
Questions used: 0/10. Admissible actions: ask / begin_workup / final_diagnosis.

Doctor Agent (Step 1)

<reasoning>
The next highest-yield information is whether the testicular mass is painful or tender and whether it is associated with scrotal swelling, which helps distinguish epididymitis/acute infection from chronic testicular torsion or tumor and guides urgency.
</reasoning>
<action>ask: Is the mass painful or tender, and do you have any scrotal swelling?</action>

Environment (Step 2)

Patient: The mass in my right testicle is not painful or tender. I haven't noticed any scrotal swelling either.
Questions used: 1/10.

Doctor Agent (Step 2)

<reasoning>
Highest-priority missing information is whether the testis is still in the scrotum versus whether it has ascended into the inguinal canal or abdomen — that determines whether this is a testicular mass versus an inguinal hernia or testicular torsion and changes urgency of evaluation and next steps.
</reasoning>
<action>ask: Is the mass in the scrotum, or has it ascended into the groin or abdomen?</action>

Environment (Step 3)

Patient: The mass is in the scrotum. It has not ascended into the groin or abdomen.
Questions used: 2/10.

Doctor Agent (Step 3)

<reasoning>
Highest-priority missing information is whether the testis is firm, soft, or normal in consistency — this distinguishes a testicular tumor (often hard/fixed) from a cyst, epididymal lesion, or varicocele (usually softer/fluctuant). If it's hard and fixed → higher concern for malignancy; if soft/fluctuant → more likely benign cyst or epididymal mass.
</reasoning>
<action>ask: Is the mass hard, soft, or normal in consistency?</action>

H OSCE STYLE FORMAT

Environment (Step 4)

Patient: The mass in my right testicle feels hard.
Questions used: 3/10.

Doctor Agent (Step 4)

<reasoning>
Highest-priority missing information is whether the testis itself is normal versus abnormal (eg, testicular atrophy, enlargement, or retraction). If the testis is normal-sized and retracted, that raises suspicion for an epididymal or spermatic cord mass (epididymitis or spermatocele); if it is enlarged or atrophic, that shifts the differential toward testicular tumor or testicular torsion/chronic torsion. I'll therefore ask about testicular size/changes.
</reasoning>
<action>ask: Has the testis changed in size or become enlarged or shrunken, and does it retract toward the groin?</action>

Environment (Step 5)

Patient: The right testicle with the mass has become enlarged, and it does not retract toward the groin.
Questions used: 4/10.

Doctor Agent (Step 5)

<reasoning>
Highest-priority missing information is whether there are associated scrotal symptoms that narrow the differential (testicular torsion, infection, varicocele, or malignancy). I should ask about scrotal pain or redness, testicular discharge, and scrotal heaviness/weight. If any of these are present, urgency or different management would change; if none, I'll proceed to next HPI elements (prior evaluation/treatment).
</reasoning>
<action>ask: Do you have any scrotal pain or redness, any discharge from the testis or scrotum, or any scrotal heaviness or a feeling of weight?</action>

Environment (Step 6)

Patient: I do not have any scrotal pain or redness, nor any discharge from the testis or scrotum. I also haven't noticed any scrotal heaviness or a feeling of weight.
Questions used: 5/10.

Doctor Agent (Step 6)

<reasoning>
Highest-priority missing information is whether the mass has been evaluated (scrotal ultrasound and tumor markers) and whether it has been treated — those results would markedly shift the differential (benign epididymal cyst or varicocele vs malignant testicular tumor). If ultrasound shows a solid hypoechoic lesion or tumor markers are abnormal, urgency for oncology referral increases; if ultrasound shows a normal testis or epididymis or markers are normal, conservative management or further evaluation is more likely. I'll therefore ask about prior evaluation and treatment.
</reasoning>
<action>ask: Have you seen a doctor for this, had a scrotal ultrasound, or been given any treatment such as antibiotics, pain medicine, or surgery for the mass?</action>

Environment (Step 7)

Patient: I have not seen a doctor for this mass yet. I have not had a scrotal ultrasound or received any treatment like antibiotics, pain medicine, or surgery for it.
Questions used: 6/10.

Doctor Agent (Step 7)

<reasoning>
Highest-priority missing information is whether the mass has changed in size over the three months and whether there are any testicular symptoms (change in testicular size, heaviness, or scrotal pain). Those details shift the differential between testicular carcinoma, benign cyst/testicular tumor, or chronic epididymitis/torsion and guide urgency for imaging or referral.
</reasoning>
<action>ask: Has the mass changed in size over the three months — gotten bigger, smaller, or stayed the same — and do you have any change in the size of your right testicle or any scrotal heaviness or pain?</action>

Environment (Step 8)

Patient: The mass in my right testicle has gotten bigger over the three months. The right testicle with the mass has become enlarged. I have not noticed any scrotal heaviness or pain.
Questions used: 7/10.

Doctor Agent (Step 8)

<reasoning>
Highest-priority missing information is whether there are systemic signs of malignancy or infection (fever, night sweats, unexplained weight loss, or new testicular pain) and whether you've had prior testicular or scrotal surgery or radiation. If you had fever or night sweats or weight loss, that would shift the differential toward infection or metastatic disease; if you had prior scrotal surgery or radiation, that could affect treatment options. I'll therefore ask about those items now.
</reasoning>
<action>ask: Have you had any fever, night sweats, or unexplained weight loss, and have you ever had testicular or scrotal surgery or radiation therapy?</action>

Environment (Step 9)

Patient: I have not had any fever, night sweats, or unexplained weight loss. I have never had testicular or scrotal surgery or radiation therapy.
Questions used: 8/10.

Doctor Agent (Step 9)

<reasoning>
Highest-priority missing information is whether the mass is associated with a testicular or scrotal cancer history (past cancers or prior testicular biopsy) and whether there is a family history of testicular cancer. A family history or prior cancer/biopsy would raise suspicion for malignancy; absence lowers but does not exclude malignancy. I will therefore ask about prior cancers/biopsies and family history.
</reasoning>
<action>ask: Do you have a family history of testicular or other cancers, and have you ever had a cancer or any testicular biopsy or other scrotal surgery?</action>

Environment (Step 10)

Patient: I don't know about any family history of testicular or other cancers. I have never had cancer myself, nor have I had any testicular biopsy or other scrotal surgery.
Questions used: 9/10.

Doctor Agent (Step 10)

<reasoning>
Highest-priority missing information is whether the patient has risk factors for testicular cancer or other testicular disease: prior sexually transmitted infections (eg, gonorrhea/chlamydia), prior undescended testis or inguinal hernia, and current tobacco/alcohol use. These factors affect urgency of evaluation (eg, scrotal ultrasound, tumor markers, possible referral) and the differential (testicular germ cell tumor vs chronic epididymitis/abscess vs other masses). I will therefore ask about those risk factors.
</reasoning>
<action>ask: Have you ever had a sexually transmitted infection such as gonorrhea or chlamydia, did you ever have an undescended testicle or an inguinal hernia, and do you currently smoke or drink alcohol?</action>

Environment (Step 11)

Patient: I don't remember being told that I ever had a sexually transmitted infection like gonorrhea or chlamydia. I also don't know if I ever had an undescended testicle or an inguinal hernia. I'm not sure about my current smoking or alcohol use.
(Topics the patient could not answer, final: sexually transmitted infection; undescended testicle; inguinal hernia; smoking; alcohol use)
Questions used: 10/10.

Doctor Agent (Step 11)

<reasoning>
Highest-priority missing information is whether the testis is still viable (i.e., whether there is scrotal erythema, testicular atrophy, or loss of normal transillumination). If the testis is atrophic or non-transilluminant, testicular torsion is unlikely and the differential shifts toward testicular tumor or chronic epididymitis. If the testis is normal in size, erythematous, and transilluminant, acute torsion is less likely but chronic torsion or epididymitis remains possible. Because the patient has a hard, enlarging scrotal mass without pain or fever and has not seen a doctor or had imaging, the next step is to get an urgent scrotal ultrasound and possibly alpha-fetoprotein/β-hCG/β-hCG (if tumor suspected). I will therefore begin a targeted workup now.
</reasoning>
<action>begin_workup</action>

Environment (Step 12)

Interview closed. The examination and test menus follow next turn.
Workup actions remaining: 20/20.

Doctor Agent (Step 12)

<reasoning>
Leading diagnoses:
- Testicular tumor (germ cell tumor, most common) — presents with a firm, non-tender, enlarging unilateral testicular mass.
- Chronic epididymitis or epididymo-orchitis — can produce a scrotal/inguinal mass and suprapublic or inguinal tenderness.
- Testicular torsion — acute, painful torsion causes scrotal swelling, severe pain, and usually erythema/erythema of scrotum; torsion is a must-not-miss condition but history is 3 months so torsion less likely.
Key alternatives to exclude urgently:
- Testicular torsion (would mandate immediate ultrasound and possible detorsion).
- Infection or abscess (would prompt antibiotics/antivirals and imaging).
- Primary testicular malignancy (would change management to oncology and systemic staging).

Which single next action will most change management:
- A careful physical exam including vital signs, general exam, scrotal (palpation for firmness, fluctuation, tenderness, transillumination, erythema), and abdominal exam. Findings that support malignancy (firm, non-tender, non-tender, transilluminant, no suprapublic tenderness) will prompt immediate scrotal ultrasound and tumor markers. Findings that suggest acute torsion (acute severe pain, scrotal erythema, high scrotal swelling, absent transillumination) would mandate emergent ultrasound and urgent urology/surgery. Suprapublic tenderness or inguinal mass would shift differential toward epididymitis or inguinal hernia and prompt different next steps.

Therefore the highest-yield next step is a focused physical exam: vital signs, general, scrotal, and abdominal exam.
</reasoning>
<action>physical_examination: vital signs, general, genitourinary, abdomen</action>

Environment (Step 13)

Results for patient PT-17244 (27-year-old male):
vital signs: Not available in record.
general: Not available in record.
genitourinary — Right testis: The right testis was normal in size and shape.
genitourinary — Lower pole of the right testis: A firm 1 cm nodule was palpable.
abdomen — Bilateral inguinal regions: Multiple elastic, non-tender nodules were palpable.
Workup actions remaining: 19.

Doctor Agent (Step 13)

<reasoning>
Leading diagnoses:
- Testicular malignancy (testicular germ cell tumor — e.g., seminoma or non-seminatous germ cell tumor): firm 1 cm nodule in the lower pole of the right testis in a 27-year-old with a 3-month history is most concerning for a testicular tumor.
- Benign testicular cyst or epididymal cyst/tumor: possible but less likely given firmness and firmness on palpation.
- Chronic epididymitis or epididymal cyst: usually tender and associated with inflammation; history lacking.
- Lymphoma or metastatic disease: less likely but possible.
Key alternatives that would change management: benign cyst/tumor versus malignancy (requires tissue diagnosis).

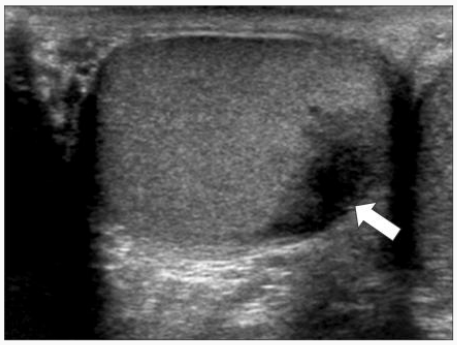
Which single next action gives the highest yield now:
- A scrotal ultrasound (Imaging: Ultrasound) to distinguish solid vs cystic lesion, assess testicular parenchyma, detect intratesticular vs extratesticular focus, and guide whether an incisional ultrasound-guided core biopsy is indicated. This will also help decide whether to proceed to CT chest/abdomen/pelvis for staging if solid lesion is present.
- Baseline labs (CBC, basic chemistries) are appropriate but are lower immediate yield than ultrasound for establishing whether biopsy is needed.

Therefore the next single action should be scrotal ultrasound plus baseline labs (CBC/chemistry) to confirm lesion character and guide urgency and type of biopsy/staging.
</reasoning>
<action>order_tests: Imaging/Ultrasound, Laboratory/Hematology</action>

Environment (Step 14)

Results for patient PT-17244 (27-year-old male):
Imaging/Ultrasound — Scrotal ultrasound: [image IMG-1 attached — integrity check next; once it passes, consult the matching specialist for the written interpretation]
[...]
Laboratory/Hematology: Not available in record.
Workup actions remaining: 18.

EVIDENCE INTEGRITY CHECK -- verify before anything else.
Your patient: PT-17244 (27-year-old male)
You ordered: Imaging/Ultrasound, Laboratory/Hematology
Returned (each served panel crop is attached to the prompt as an image):
IMG-1: PMC3085626_img_1#FIG1
Reply with a <reasoning> block (modality and anatomy of EACH image, read from the image alone; compare with the order, the patient's identity/demographics and the history), then exactly one action:
<action>gate_verdict: pass</action> | <action>gate_verdict: mismatch -- assets: ...; axes: patient|modality|anatomy</action>



IMG-1

Doctor Agent (Step 14)

<reasoning>
IMG-1 — Modality and anatomical site (from image alone):
- Modality: Ultrasound (grayscale transabdominal/transperineal-type image).
[...]
Comparison to order, patient identity/demographics, and history:
- Ordered studies included Imaging/Ultrasound and Laboratory/Hematology. IMG-1 is an ultrasound study of the scrotum/testis, matching the ordered imaging.
- Patient identity/demographics: No printed identifiers are present on the image (normal for a de-identified journal-style image). There is no positive conflict with the documented patient (PT-17244, 27-year-old male).
- Anatomy: The image shows the scrotum/testis as ordered; no conflicting anatomy is present.
- Conclusion: The returned image is an ultrasound of the scrotum/testis and is consistent with the ordered ultrasound for this patient. No mismatches detected.
</reasoning>
<action>gate_verdict: pass</action>

Environment (Step 15)

verdict recorded.

Doctor Agent (Step 15)

<reasoning>
Leading diagnoses:
- Testicular tumor (testicular germ cell tumor or less likely primary testicular lymphoma) — firm 1 cm nodule at lower pole of right testis, 3 months duration, normal-appearing contralateral testis.
- Local cyst or epididymal lesion (epididymitis or epididymal cyst) — ultrasound may distinguish solid vs cystic.
- Lymphoma or metastatic lesion to scrotum (rare).
Key alternatives that would change management: testicular malignancy (would mandate scrotal ultrasound read, tumor markers, scrotal MRI/CT chest/abdomen, and prompt oncology referral), benign cyst/epididymal process (less urgent but still need characterization).

What single next action will most change management now:
- Obtain an expert interpretation of the already-obtained scrotal ultrasound (IMG-1). A radiologist read will confirm whether the lesion is solid vs cystic, whether it involves the testis or epididymis, and whether there are suspicious features (e.g., solid hypoechoic nodule, calcification, heterogeneous internal echoes) that would raise suspicion for malignancy and guide the next steps (tumor markers, scrotal MRI, CT chest/abdomen, Doppler, biopsy). Integrity check already passed for IMG-1.

Therefore the highest-yield next step is to request a radiology consult for the scrotal ultrasound.
</reasoning>
<action>consult_specialist: radiology</action>

Environment (Step 16)

radiology interpretation (patient PT-17244):
- Scrotal ultrasound (Ultrasound): An irregular, poorly defined hypoechoic lesion was present at the lower pole of the right testis, with increased surrounding blood flow and parenchymal distortion.
Workup actions remaining: 17.

Doctor Agent (Step 16)

<reasoning>
Leading diagnosis: intratubular germ cell neoplasm of the testis (ITGCN) — a germ cell neoplasm that presents as a small, firm intratubular nodule with normal-appearing external testis on exam and ultrasound. Key alternatives that would change management: seminoma or nonseminatous germ cell tumor (would typically be more solid, heterogeneous, or associated with an epididymal tail), intratubular neoplasm progressing to invasive germ cell tumor, or an inflammatory/infectious lesion (rare in this age group).

Why the differential matters: ITGCN is an early germ cell neoplasm that, when detected early, is associated with better long-term prognosis and different staging/therapy (often local excision or orchiectomy with careful follow-up). If invasive disease or a different tumor type is present, management shifts to systemic therapy and broader staging (CT chest/abdomen/pelvis, tumor markers).

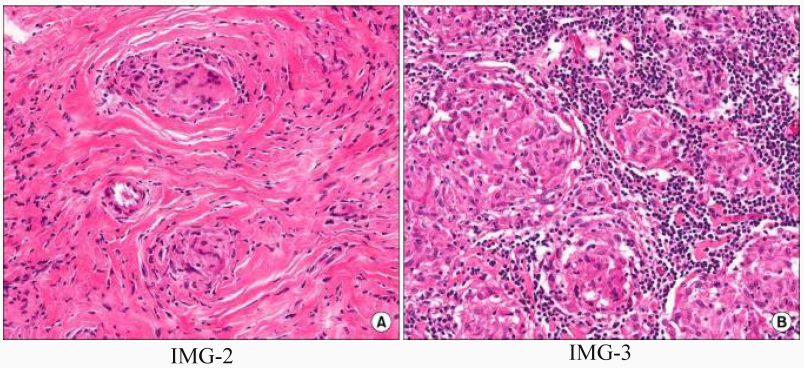
What single next action will most change management now:
- Tissue diagnosis. The scrotal ultrasound shows a small, poorly defined hypoechoic lesion with increased flow and parenchymal distortion — features that could be consistent with ITGCN or early invasive germ cell tumor. The highest-yield next step is to obtain tissue (biopsy or core/needle biopsy) so histology can distinguish ITGCN from seminoma/nonseminatous germ cell tumors and guide definitive therapy.

Therefore the next immediate action should be to request histopathologic evaluation (biopsy) of the lesion, plus targeted tumor markers (AFP, β-hCG, LDH) to support germ cell neoplasm diagnosis and stage risk.
</reasoning>
<action>order_tests: Pathology/Histopathology, Laboratory/Tumor Markers</action>

Environment (Step 17)

Results for patient PT-17244 (27-year-old male):
Pathology/Histopathology — Right testicular lesion histopathology: [image IMG-2 attached — integrity check next;
[...]
Workup actions remaining: 16.

EVIDENCE INTEGRITY CHECK -- verify before anything else.
Your patient: PT-17244 (27-year-old male)
You ordered: Pathology/Histopathology, Laboratory/Tumor Markers
Returned (each served panel crop is attached to the prompt as an image):
IMG-2: PMC3085626_img_3#FIG3A
IMG-3: PMC3085626_img_3#FIG3B
Reply with a <reasoning> block (modality and anatomy of EACH image, read from the image alone; compare with the order, the patient's identity/demographics and the history), then exactly one action:
<action>gate_verdict: pass</action> | <action>gate_verdict: mismatch -- assets: ...; axes: patient|modality|anatomy</action>



IMG-2 IMG-3

Doctor Agent (Step 17)

<reasoning>
- Modality (from image): brightfield light microscopy histopathology (H&E stain). [...]
- Comparison to order/history:
  - Ordered category: Pathology/Histopathology — MATCH.
  - Patient identity/demographics: No printed identifiers on image (normal for de-identified journal figures). No positive conflict with PT-17244 (27-year-old male).
  - Anatomy: Consistent with the requested right testicular lesion histopathology. [...]
Overall assessment:
- Both images are histopathology slides and match the ordered Pathology/Histopathology studies.
- Images show the correct anatomical sites (testis and lymph node) and show no conflicting patient identifiers or demographics.
- No positive mismatch detected.
</reasoning>
<action>gate_verdict: pass</action>

Environment (Step 18)

verdict recorded.

Doctor Agent (Step 18)

<reasoning>
Leading diagnosis: intratubular germ cell neoplasia in situ (teratoma with associated intratubular germ cell neoplasia) — this fits a young man with a 3-month firm testicular nodule, scrotal ultrasound showing a hypoechoic lower-pole lesion with increased blood flow and parenchymal distortion, and an incidental normal left testis. Key alternatives: (1) seminoma or nonseminatous germ cell tumor (less likely given normal AFP/beta-hCG/LDH and normal left testis); (2) testicular lymphoma or metastatic lesion (less likely but possible); (3) benign cystic lesion or intratubular germ cell neoplasia without a macroscopic mass (would be seen on histology).

What would most change management/diagnosis now: definitive histologic interpretation of the available pathology images (IMG-2 and IMG-3). Histology can distinguish germ cell neoplasia (including intratubular germ cell neoplasia), seminoma, nonseminatous elements, lymphoma, and benign processes — and it will guide whether further oncologic staging (CT chest/abdomen/pelvis, additional imaging, systemic therapy) is indicated.

Therefore the single best next action is to obtain an expert pathology read of the submitted histopathology images.
</reasoning>
<action>consult_specialist: pathology</action>

Environment (Step 19)

pathology interpretation (patient PT-17244):
- Right testicular lesion histopathology (Histopathology): Non-caseating epithelioid cell granulomas with giant cells were present.
- Right superficial inguinal node histopathology (Histopathology): Non-caseating epithelioid cell granulomas with giant cells were present.
Workup actions remaining: 15.

Doctor Agent (Step 19)

<reasoning>
Leading diagnosis: granulomatous testicular tuberculosis (or disseminated mycobacterial infection) causing a focal testicular lesion with regional lymph node involvement.

Key alternatives:
- Noncaseating mycobacterial infection (eg, atypical mycobacteria) mimicking TB.
- Granulomatous fungal infection (eg, histoplasmosis, cryptococcus).
- Granulomatous sarcoidosis (less likely to present with a discrete palpable scrotal mass).
- Non-tuberculous mycobacterial or fungal dissemination from an extragenital focus (eg, pulmonary TB or abdominal focus).

What would most change management/diagnosis now:
- Microbiologic confirmation (AFB and fungal studies) on the histopathology specimens and on any available tissue/biopsy. This distinguishes TB from noncaseating mycobacteria or fungal infection and guides systemic therapy.
- Cross-sectional imaging to evaluate for extragenital foci (eg, pulmonary TB, abdominal/inguinal lesions) that could explain dissemination.

Given the available histopathology (noncaseating granulomas with giant cells) and the safety/management implications of treating TB vs. other granulomatous infections, the single highest-yield next step is to request microbiologic testing on the existing histopathology slides and to obtain CT chest/abdomen/pelvis to look for extragenital sites of infection.

Therefore order targeted microbiology studies (AFB and fungal stains/cultures on histology) and a CT chest/abdomen/pelvis.
</reasoning>
<action>order_tests: Imaging/Computed Tomography, Laboratory/Microbiology</action>

Environment (Step 20)

Results for patient PT-17244 (27-year-old male):
Imaging/Computed Tomography — Abdominal CT scan: Extensive lymph node enlargement was present in the cardiophrenic, mediastinal, retroperitoneal, pelvic, and inguinal regions. [...]
Laboratory/Microbiology: Not available in record.
Workup actions remaining: 14.

EVIDENCE INTEGRITY CHECK -- verify before anything else.
Your patient: PT-17244 (27-year-old male)
You ordered: Imaging/Computed Tomography, Laboratory/Microbiology
Returned (each served panel crop is attached to the prompt as an image):
IMG-4: PMC3085626_img_2#FIG2A
IMG-5: PMC3085626_img_2#FIG2B
Reply with a <reasoning> block (modality and anatomy of EACH image, read from the image alone; compare with the order, the patient's identity/demographics and the history), then exactly one action:
<action>gate_verdict: pass</action> | <action>gate_verdict: mismatch -- assets: ...; axes: patient|modality|anatomy</action>



IMG-4 IMG-5

Doctor Agent (Step 20)

<reasoning>

- Modality (from image appearance): Contrast-enhanced CT (axial slices).
- Anatomical site (from image appearance): Abdomen (upper abdomen showing liver, kidneys, spleen, pancreas, stomach, and pelvis/iliopsoas).
- Comparison to order and patient info:
  - Ordered: Imaging/Computed Tomography — matches (contrast-enhanced CT).
  - Anatomical site: Abdomen — matches the ordered abdominal CT.

[...]
Conclusion:
- Both returned images are contrast-enhanced abdominal CTs and match the ordered CT study for the patient. No patient, modality, or anatomy mismatches are present.
</reasoning>
<action>gate_verdict: pass</action>

Environment (Step 21)

verdict recorded.

Doctor Agent (Step 21)

<reasoning>
Leading diagnosis: sarcoidosis with granulomatous testicular involvement and extratesticular (inguinal, mediastinal, retroperitoneal, pelvic) lymphadenopathy — supported by non-caseating epithelioid cell granulomas with giant cells in testis and superficial inguinal node, and extensive lymphadenopathy on CT.

Key alternatives:
- Granulomatous infection (e.g., Mycobacterium tuberculosis, fungal infection) — can produce non-caseating granulomas and systemic lymphadenopathy.
- Lymphoma with granulocytic reaction (less likely but possible).
- Primary testicular malignancy with nodal metastasis (less likely given histology).

What would change management/diagnosis most right now:
- Identification of infectious etiology (TB or fungal) or evidence of active extranodal disease/complications. Tissue diagnosis (AFB/fungal stains, molecular testing on biopsy specimens) and targeted immunologic markers (ACE, IBD markers) would be highest yield. However, the immediate safety/management concern is whether there is active disseminated infection (e.g., tuberculous peritonitis/abscess) or extranodal disease that could alter therapy (immunosuppression vs antimicrobial therapy).

Given the CT shows extensive nodal disease with peritoneal involvement and the patient was already on high-dose corticosteroids, the single next action that will change management now is to get expert radiology interpretation of the contrast CT images (IMG-4 and IMG-5) to distinguish active infection/abscess versus fibrosis or lymphoma, assess for extranodal spread (pleural, peritoneal, hepatic), and guide need for targeted biopsy or further imaging.

Therefore request radiologist read of the contrast CTs (IMG-4, IMG-5) now.
</reasoning>
<action>consult_specialist: radiology</action>

Environment (Step 22)

radiology interpretation (patient PT-17244):
- Contrast-enhanced abdominal CT (Computed Tomography): Extensive enlarged lymph nodes were present in the retroperitoneal space.
- Contrast-enhanced abdominal CT (post-treatment — Four months after the start of high-dose corticosteroid therapy) (Computed Tomography): Multiple extensive retroperitoneal lymph nodes had considerably decreased in size.
Workup actions remaining: 13.

Doctor Agent (Step 22)

<reasoning>
Leading diagnosis: sarcoidosis with granulomatous testicular involvement and extratesticular (retroperitoneal) lymphadenopathy. Rationale: biopsy of the testicular lesion and an inguinal node both showed non-caseating epithelioid-cell granulomas with giant cells — classic for sarcoidosis. CT shows extensive nodal disease and omental/peritoneal nodules that regressed after high-dose corticosteroids, supporting a systemic granulomatous inflammatory process (sarcoidosis) rather than an infectious or neoplasm. Testicular tumor markers were normal, making germ cell tumor unlikely.

Key alternatives to exclude:
- Granulomatous infection (e.g., tuberculosis, fungal) — can cause non-caseating granulomas.
- Lymphoma or metastatic malignancy with granulomatous reaction — less likely but possible.
- Sarcoidosis with concurrent infection (e.g., TB) — important to rule out because it changes management (anti-TB therapy).

Which single next action will most change management/diagnosis now:
- The highest-yield next step is targeted laboratory testing to exclude infectious mimics (TB and fungal infection) and to screen for other systemic causes. Specifically, molecular testing for Mycobacterium tuberculosis (PCR) and fungal PCR on the biopsy specimens (molecular genetics) will rapidly distinguish infection from sarcoidosis. Simultaneously, immunologic testing (e.g., immunoglobulins, ACE if available) and endocrine testing (testosterone, other relevant hormones) can help characterize systemic activity and exclude alternative diagnoses. These tests are the most urgent and will guide whether further immunosuppression or antimicrobial therapy is indicated.

Therefore order targeted molecular studies and immunologic/endocrine labs.
</reasoning>
<action>order_tests: Laboratory/Molecular Genetics, Laboratory/Immunology, Laboratory/Endocrinology</action>

Environment (Step 23)

Results for patient PT-17244 (27-year-old male):
Laboratory/Molecular Genetics: Not available in record.
Laboratory/Immunology: Not available in record.
Laboratory/Endocrinology: Not available in record.
Workup actions remaining: 12.

Doctor Agent (Step 23)

<reasoning>
Immunology was already placed and cannot be re-requested. The next best single action that will materially change management and distinguish sarcoidosis from infection is to obtain targeted molecular testing (eg, PCR for mycobacterial DNA and fungal PCR on tissue/fluid) and to assess immunosuppression/safety labs (eg, CBC, CMP, coagulation) to guide steroid therapy. These tests are admissible and will change management.
</reasoning>
<action>order_tests: Laboratory/Coagulation</action>

Environment (Step 24)

Results for patient PT-17244 (27-year-old male):
Laboratory/Coagulation: Not available in record.
[Laboratory/Molecular Genetics: already ordered, skipped]
Workup actions remaining: 11.

Doctor Agent (Step 24)

<reasoning>
Leading diagnosis: sarcoidosis with granulomatous involvement of testis and inguinal node and secondary extensive nodal disease. Rationale: firm testicular nodule with non-caseating epithelioid cell granulomas on histology, systemic extensive lymphadenopathy on CT, and partial regression of nodes after high-dose steroids — all most consistent with sarcoidosis.

Key alternatives:
- Granulomatous infection (e.g., Mycobacterium tuberculosis, atypical mycobacteria, deep fungal infection) — can produce non-caseating granulomas and lymphadenopathy; would be supported by microbiology or PCR results (not available).
- Granulomatous malignancy (e.g., lymphoma with granulocytic reaction, rare metastatic disease) — less likely but possible; would show malignant cells on histology.
- Foreign-body or drug-induced granulomatous reaction — less likely without history of surgery or foreign-body exposure.

What single next action will most change management/diagnosis now:
- A tissue diagnosis from the CT-identified peritoneal/retroperitoneal nodules (biopsy) is essential to exclude infection or malignancy and to guide therapy. Histology/cytology from those lesions will distinguish sarcoidosis from infection or neoplasm and help explain the regression seen on post-steroid CT.

Safety/systems to cover: genitourinary (scrotum/inguinal nodes), abdominal (retroperitoneal/peritoneal disease), respiratory (mediastinal nodes) — all touched by the differential and treatment.

Therefore the next best single action is to obtain pathology from the peritoneal/retroperitoneal nodules (image-guided biopsy) so histopathology/cytology can be performed and a definitive diagnosis made.
</reasoning>
<action>order_tests: Pathology/Bone Marrow Examination</action>

Environment (Step 25)

Results for patient PT-17244 (27-year-old male):
Pathology/Bone Marrow Examination: Not available in record.
Workup actions remaining: 10.

Doctor Agent (Step 25)

<reasoning>
Leading diagnosis: sarcoidosis with granulomatous testicular involvement and extragenital sarcoidosis (mediastinal, retroperitoneal, pelvic lymphadenopathy and peritoneal nodules).
Rationale: firm testicular nodule with non-caseating epithelioid cell granulomas on histology; extensive nodal disease and peritoneal nodules on CT; partial regression of nodes after high-dose steroids — all classic for sarcoidosis.

Key alternatives that would change management:
- Granulomatous infection (eg, tuberculosis): can mimic sarcoidosis histologically; would require targeted microbiology and possibly antituberculous therapy.
- Lymphoma or other malignancy with granulomatous reaction: less likely but possible; would be suggested by atypical cytology/flow cytometry.
- Primary testicular neoplasm with granulomatous inflammation: unlikely given granuloma-predominant histology and tumor marker normality.

What single action will most change management now:
- A tissue diagnosis from the peritoneal/retroperitoneal lesion (cytopathology/cytology) to distinguish sarcoidosis from infection or malignancy and to guide therapy. The prior histopathology was from testis/biopsies; cytologic evaluation of peritoneal fluid or node aspirate/cytology will directly change the differential and immediate treatment (eg, antituberculous therapy if TB found, or adjust immunosuppression).

Therefore the next highest-yield step is to obtain cytopathology on available peritoneal/retroperitoneal tissue (eg, peritoneal fluid cytology or node aspirate cytology).
</reasoning>
<action>order_tests: Pathology/Cytopathology</action>

Environment (Step 26)

Results for patient PT-17244 (27-year-old male):
Pathology/Cytopathology: Not available in record.
Workup actions remaining: 9.

Doctor Agent (Step 26)

<reasoning>
Leading diagnosis: sarcoidosis with granulomatous testicular involvement and reactive (non-caseating) granulomatous lymphadenopathy. Supporting data: firm 1 cm hypoechoic lower-pole testicular nodule on scrotal US with parenchymal distortion and increased blood flow; histopathology showing non-caseating epithelioid-cell granulomas with giant cells in the right testis and an adjacent superficial inguinal node; extensive CT-confirmed lymphadenopathy (retroperitoneal, mediastinal, inguinal, pelvic) that markedly regressed with high-dose corticosteroids. Normal tumor markers (AFP, beta-hCG, LDH) make testicular germ cell tumor less likely. Testicular lymphoma or tuberculosis is a key alternative, but non-caseating granulomas and marked steroid responsiveness favor sarcoidosis.

Key alternatives to exclude:
- Testicular tuberculosis (would show caseating granulomas or organisms on special stains/culture).
- Granulomatous lymphoma (would show atypical lymphoid proliferation; bone marrow pathology was ordered but results not available).
- Infectious or inflammatory mimics (e.g., fungal infection).

What single next action will most change management/diagnosis now:
- Obtain tissue diagnosis from the lesion to confirm non-caseating granulomatous inflammation (sarcoidosis) versus caseating organisms or lymphoma. Histology alone was reported; targeted special stains and molecular studies on the existing biopsy/tissue (AFB, fungal stains, PCR for Mycobacterium tuberculosis and relevant pathogens, and molecular markers) are the highest-yield next step to exclude tuberculosis and other mimics and to support sarcoidosis as the working diagnosis. This will directly change immunosuppressive vs antimicrobial therapy.

Therefore the next action is to request pathology special stains and molecular testing on the available biopsy specimens.
</reasoning>
<action>order_tests: Pathology/Special Stains and Histochemistry, Pathology/Molecular Genetics</action>

Environment (Step 27)

Results for patient PT-17244 (27-year-old male):
Pathology/Special Stains and Histochemistry: Not available in record.
Pathology/Molecular Genetics: Not available in record.
Workup actions remaining: 8.

Doctor Agent (Step 27)

<reasoning>
Leading diagnosis: sarcoidosis with granulomatous testicular involvement and reactive (non-caseating) granulomatous lymphadenopathy. Rationale: biopsy histology shows non-caseating epithelioid cell granulomas in testis and inguinal node; CT shows extensive lymphadenopathy that decreased with high-dose steroids — classic for sarcoidosis. Testicular tumor markers were normal, arguing against a germ cell neoplasm.

Key alternatives:
- Granulomatous infection (eg, Mycobacterium tuberculosis or other mycobacteria, fungal infection) — can produce non-caseating granulomas and mimic sarcoidosis. Important because therapy differs (antimycobacterial/antifungal vs immunosuppressive).
- Leprosy (Mycobacterium leprae) — can cause non-caseating granulomas and testicular involvement; less common but relevant.
- Foreign-body granuloma or Crohn disease (less likely).

Which single next action will most change management/diagnosis now:
- The histopathology images and radiology reports are present, but the pathology slides lack microbiologic studies (AFB, PCR, fungal stains, etc.) and molecular testing was ordered but returned Not available. The highest-yield next step is to search the literature for the phenotype (non-caseating epithelioid cell granulomas in testis and node with extensive lymphadenopathy that regress with corticosteroids) to confirm the differential and guide whether additional targeted microbiology/molecular tests are required urgently.

Therefore I will perform a targeted literature search on that phenotype.
</reasoning>
<action>search: What conditions cause non-caseating epithelioid cell granulomas in the testis and lymph nodes with extensive lymphadenopathy that regress with corticosteroid therapy?</action>

Environment (Step 28)

<information>
[E1 | PubMed 9168651] [Systemic sarcoidosis with involvement of the spermatic cord].
A 49-year-old man presented with an enlarging, painful, right inguinal mass and small red nodules around his nose and upper lip. He had undergone an operation for a right inguinal hernia in childhood. Because malignancy was suspected, the right spermatic cord, testis, and epididymis were resected. [...]
URL: https://pubmed.ncbi.nlm.nih.gov/9168651/

[E2 | PubMed 35381637] [Subcutaneous nodules, mediastinal and hilar lymphadenopathy]. A 52-year old man was admitted to our hospital because of dyspnea on exertion for 2 months and subcutaneous nodules for 1 month. [...]
The patient was diagnosed as sarcoidosis and glucocorticoid was administrated. The subcutaneous nodules were improved.
URL: https://pubmed.ncbi.nlm.nih.gov/35381637/

[E3 | PubMed 19926953] [Epididymal sarcoidosis: a case report].
A 33-year-old male complained of poor vision and visited the ophthalmology department of our hospital. Right iridoncus suspected as a metastatic nodule, was found. [...]
A hard and nodular mass at the head of the epididymis, completely separated from the testis, was recognized. Only open biopsy was done, because no malignancy was found by the intraoperative biopsy.
URL: https://pubmed.ncbi.nlm.nih.gov/19926953/
</information>
Workup actions remaining: 7.

Doctor Agent (Step 28)

<reasoning>
Leading diagnosis: sarcoidosis with testicular involvement — supported by non-caseating epithelioid-cell granulomas in testis and inguinal node, extensive lymphadenopathy, and regression of nodes with corticosteroids. Key alternatives (less likely): granulomatous infection (e.g., TB) or testicular malignancy.

What will most change management now:
- Identifying whether this exact pattern (testicular non-caseating granulomas + extensive lymphadenopathy + steroid-responsive regression) is well documented will confirm sarcoidosis and guide further testing (e.g., search for extragenital disease, systemic immunosuppression, and need for microbiology/special stains). A literature search for that specific clinical picture is the highest-yield next step.

Refined search to capture the exact pattern:
- Include terms for non-caseating granulomas, steroid-responsive regression, testis involvement, and lymphadenopathy to ensure we capture cases matching this presentation.
</reasoning>
<action>search: What conditions cause non-caseating epithelioid cell granulomas in the testis and lymph nodes, extensive lymphadenopathy, and regression of lesions with corticosteroid therapy?</action>

Environment (Step 29)

<information>
[E1 | PubMed 19926953] [Epididymal sarcoidosis: a case report].
A 33-year-old male complained of poor vision and visited the ophthalmology department of our hospital. Right iridoncus suspected as a metastatic nodule, was found. A systemic examination was done. Abdominal computed tomographic (CT) scan showed an enhanced mass in the right scrotum, 15 mm in diameter and enlargement of iliac lymph nodes. Chest CT scan demonstrated bilateral enlargement of mediastinal lymph nodes and enlargement of lymph nodes above the collarbone. Since malignant lymphoma or testicular malignancy was suspected, right orchiectomy was planned. A hard and nodular mass at the head of the epididymis, completely separated from the testis, was recognized. Only open biopsy was done, because no malignancy was found by the intraoperative biopsy. Pathological findings of the epididym
URL: https://pubmed.ncbi.nlm.nih.gov/19926953/

[E2 | PubMed 9168651] [Systemic sarcoidosis with involvement of the spermatic cord].
A 49-year-old man presented with an enlarging, painful, right inguinal mass and small red nodules around his nose and upper lip. He had undergone an operation for a right inguinal hernia in childhood. Because malignancy was suspected, the right spermatic cord, testis, and epididymis were resected. Histological examination of the lesion in the spermatic cord revealed non-caseating epithelioid-cell granuloma and no acid-fast bacilli, which was compatible with sarcoidosis. [...]
URL: https://pubmed.ncbi.nlm.nih.gov/9168651/

[E3 | PubMed 912651] Malignant lymphoma obscured by concomitant extensive epithelioid granulomas: report of three cases with similar clinicopathologic features.
Three similar cases are described of an unusual combination of malignant lymphoma and extensive non-necrotic granulomas. The three patients presented with prominent splenomegaly without peripheral lymphadenopathy. They had normal or moderately elevated lymphocyte counts, abnormal lymphoid cells in the peripheral blood and bone marrow, and abnormalities of serum immunoglobulins. The lymphoid tumor was difficult to recognize but it was best identified in abdominal lymph nodes, it was composed of small atypical lymphocytes proliferating in a vaguely nodular pattern. The presence of multiple epithelioid granulomas obscured the neoplastic proliferation in the spleens and misled or delayed the final interpretation of the malignant disease.
URL: https://pubmed.ncbi.nlm.nih.gov/912651/
</information>
Workup actions remaining: 6.

Doctor Agent (Step 29)

<reasoning>
Pathology cannot be consulted because no new study is awaiting interpretation (histology images were already reviewed). The next best single action is to request targeted immunohistochemistry on the existing biopsy specimens (CD3, CD20, CD45, and additional markers as indicated) to exclude lymphoma or other mimics; this will materially affect management and is admissible now.
</reasoning>
<action>order_tests: Pathology/Immunohistochemistry</action>

Environment (Step 30)

Results for patient PT-17244 (27-year-old male):
Pathology/Immunohistochemistry: Not available in record.
Workup actions remaining: 5.

Doctor Agent (Step 30)

<reasoning>
Leading diagnosis: sarcoidosis with testicular and nodal involvement. Rationale: young man with firm 1-cm testicular nodule at lower pole, scrotal ultrasound showing irregular hypoechoic lesion with parenchymal distortion, histopathology of testis and inguinal node showing non-caseating epithelioid-cell granulomas with giant cells, and extensive lymphadenopathy that regressed with high-dose corticosteroids — all classic for sarcoidosis.

Key alternatives:
- Mycobacterial infection (eg, atypical mycobacteria or Mycobacterium tuberculosis) causing non-caseating granulomas — less likely because histology was non-caseating and there was regression with corticosteroids.
- Fungal infection (eg, histoplasmosis) — can cause non-caseating granulomas but typically with caseation or different clinical features.
- Granulomatous lymphoma or other neoplasm — possible but histology favors granulomatous inflammation rather than a dominant neoplasm; regression on steroids argues against malignancy.
- Other granulomatous disorders (eg, Crohn disease) — less likely given testicular and nodal involvement.

What would change management: tissue diagnosis for organisms or demonstration of systemic infection (eg, chest imaging) — but chest CT/X-ray data were not provided in the record. The single highest-yield next step that is available and will most influence diagnosis and counseling is to obtain targeted literature evidence that explicitly supports the diagnosis of testicular/spermatic cord sarcoidosis in a young man with this constellation of findings (testicular nodule, non-caseating granulomas, extensive lymphadenopathy, steroid-responsive regression). A focused literature search for this exact clinical scenario will provide authoritative support for the working diagnosis and guide further management (eg, need for systemic therapy, monitoring, and exclusion of mimics).

Therefore the next action is a targeted literature search for conditions that produce non-caseating epithelioid-cell granulomas in testis and lymph nodes with extensive lymphadenopathy and regression on corticosteroids.
</reasoning>
<action>search: What conditions cause non-caseating epithelioid-cell granulomas in the testis and lymph nodes with extensive lymphadenopathy that regress with corticosteroid therapy?</action>

Environment (Step 31)

<information>
[E1 | PubMed 35381637] [Subcutaneous nodules, mediastinal and hilar lymphadenopathy].
A 52-year old man was admitted to our hospital because of dyspnea on exertion for 2 months and subcutaneous nodules for 1 month. Chest enhanced CT showed bilateral hilar and mediastinal lymphadenopathy. [...]
The patient was diagnosed as sarcoidosis and glucocorticoid was administrated.
URL: https://pubmed.ncbi.nlm.nih.gov/35381637/

[E2 | PubMed 912651] Malignant lymphoma obscured by concomitant extensive epithelioid granulomas: report of three cases with similar clinicopathologic features. [...]
The presence of multiple epithelioid granulomas obscured the neoplastic proliferation in the spleens and misled or delayed the final interpretation of the malignant disease.
URL: https://pubmed.ncbi.nlm.nih.gov/912651/

[E3 | PubMed 33708313] Painful testicular swelling: an unusual localization of sarcoidosis.
Sarcoidosis is a systemic inflammatory disease characterized by non-caseating epitheloid granulomas; whereas it usually involves the lungs and lymph nodes. Testicular sarcoidosis is extremely rare, having been reported to occur in 0.2% of all sarcoidosis patients.
URL: https://pubmed.ncbi.nlm.nih.gov/33708313/
</information>
Workup actions remaining: 4.

Doctor Agent (Step 31)

<reasoning>
Leading diagnosis: sarcoidosis with testicular and nodal involvement. Rationale: biopsy-confirmed non-caseating epithelioid-cell granulomas in testis and inguinal node; extensive lymphadenopathy on CT; marked regression of retroperitoneal nodes after high-dose corticosteroids. This pattern is classic for systemic sarcoidosis involving the scrotum/testis.

Key alternatives:
- Granulomatous infection (e.g., Mycobacterium tuberculosis, atypical mycobacteria) — can cause non-caseating granulomas but would be expected to show caseation or organisms on special stains/cultures (results not available).
- Lymphoma with granulomatous reaction — possible but histology favors granulomatous inflammation; regression with steroids argues against malignancy.
- Iatrogenic granulomatosis (steroid-induced) — less likely given biopsy evidence.

Which single next action will best advance or confirm the diagnosis:
- Obtain evidence of extragenital sarcoidosis (lung involvement, cutaneous lesions) to establish systemic disease and support the diagnosis. The most informative next steps are chest imaging (to detect hilar/mediastinal lymphadenopathy or pulmonary infiltration) and targeted skin examination for cutaneous nodules or erythema.
[...]

Therefore order chest radiography (computed radiography) and a focused skin examination.
</reasoning>
<action>order_tests: Imaging/Computed Radiography, Physical Examination/skin</action>

Environment (Step 32)

Results for patient PT-17244 (27-year-old male):
Imaging/Computed Radiography — Chest X-ray: Normal.
Physical Examination/skin: Not available in record.
Workup actions remaining: 3.

Doctor Agent (Step 32)

<reasoning>
Leading diagnosis: sarcoidosis with testicular and nodal involvement. Rationale: biopsy-confirmed non-caseating epithelioid-cell granulomas in testis and inguinal node; extensive lymphadenopathy on CT; marked regression of retroperitoneal nodes after high-dose corticosteroids. This pattern is classic for systemic sarcoidosis involving the scrotum/testis.

Key alternatives:
- Granulomatous infection (e.g., Mycobacterium tuberculosis, atypical mycobacteria) — can cause non-caseating granulomas but would expect caseation in some lesions or positive AFB/PCR; treatment response differs.
- Granulomatous lymphoma or lymphoma masked by granulomas — possible but histology favors granulomatous inflammation.
- Iatrogenic granulomas from prior therapy or unusual autoimmune process — less likely.

What additional information would change management or the working diagnosis?
- Tissue-specific microbiologic studies (AFB, PCR for mycobacteria) and targeted immunophenotyping/flow cytometry on biopsy specimens would help exclude infection or lymphoma.
- A full immunologic evaluation (T-cell subsets, HIV) and baseline labs (CBC, CMP, LDH, calcium, ACE) are important, but many of these were ordered earlier and results are not available in the record.
Given the available histology and the documented dramatic regression after corticosteroids, the most likely and highest-yield conclusion is systemic sarcoidosis involving the testis and nodal disease.

Therefore the single best action now is to state the most likely diagnosis so treatment planning can proceed (e.g., long-term immunosuppression, chest imaging, ophthalmology evaluation).
</reasoning>
<action>final_diagnosis: Sarcoidosis</action>

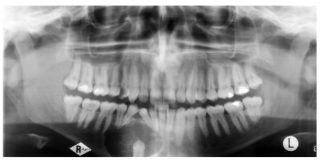
Environment (end of episode)

Episode closed by final_diagnosis. Gold diagnosis: Sarcoidosis. Judge (gpt-4.1-mini, ontology floor): full -> t_reward 1.00.

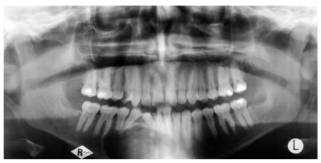
```
{
 "pmcid": "PMC1266042", "diagnosis": "Adenomatoid Odontogenic Tumor", "orpha_name": "Rare odontogenic tumor",
 "case_report": "A 23-year-old male was referred by his general dental practitioner. One year prior, an x-ray revealed a cyst associated with an ectopic lower right canine tooth. The patient had an uneventful medical history and no notable intraoral clinical findings, except for the absence of tooth 43. Radiographic examination showed a 3 cm unicystic radiolucent area with clear boundaries. Tooth 43 was situated at the base of this area. There was no evidence of root apex resorption. The lesion was surgically removed under general anesthesia and the cavity was filled with pelvic spongiosa. The separation of the lesion from the mandibular bone was uncomplicated. The postoperative recovery was uneventful.",
 "specialty": ["oral and maxillofacial surgery", "dentistry"], "image_id": ["PMC1266042_img_1", "PMC1266042_img_2"], "table_id": [], "n_images": 2, "n_panels": 2,
 "n_post_baseline": 2,
 "OSCE_Examination": {
  "Objective_for_Doctor": "Evaluate a patient referred for a previously identified radiolucent lesion in the lower right jaw associated with an unerupted tooth.",
  "Encounter_Context": {"fact_id": "CTX-01", "atomic_fact": "The patient was referred by his general dental practitioner."},
  "Patient_Actor": {
   "Demographics": {"fact_id": "DEM-01", "atomic_fact": "23-year-old male"},
   "Chief_Complaint": [],
   "History_of_Present_Illness": [
    {"fact_id": "HPI-01", "atomic_fact": "An x-ray one year earlier revealed a lesion associated with an ectopic lower right canine tooth."}
   ],
   "Denied_Symptoms": [],
   "Past_Medical_History": [
    {"fact_id": "PMH-01", "atomic_fact": "Uneventful medical history."}
   ],
   "Medication_History": [], "Surgical_or_Procedural_History": [], "Family_History": [], "Social_History": [], "Review_of_Systems": []
  }
 },
```



```
 "Physical_Examination_Findings": {
  "Oropharyngeal_Examination": [
   {"location": "Oral cavity", "findings": "No notable intraoral clinical findings were reported apart from absence of tooth 43."}
  ]
 },
 "Laboratory_Test_Results": [], "Physiologic_Test_Results": [],
 "Imaging_Results": [
  {"test_name": "Panoramic radiograph", "findings": "A 3 cm unilocular radiolucent area with clear boundaries is present in the lower right jaw. Tooth 43 is situated at the base of the area, without root apex resorption.", "panel_id": "FIG1", "modality": "xray", "source_order": 1, "view": "panoramic", "image_id": "PMC1266042_img_1"}
 ],
 "Pathology_Results": [],
 "Post_Baseline_Findings": [
  {"section": "Physical_Examination_Findings", "test_name": "Postoperative recovery", "findings": "Postoperative recovery was uneventful.", "phase": "post_treatment"},
  {"section": "Imaging_Results", "test_name": "Follow-up panoramic radiograph", "findings": "No root resorption was observed.", "phase": "follow_up", "timepoint": "six months after therapy", "panel_id": "FIG5", "modality": "xray", "source_order": 2, "image_id": "PMC1266042_img_2"}
 ],
}
```

PMC1266042_img_1



PMC1266042_img_2

Figure 9: Example of a multi-turn clinical diagnostic trajectory in our interactive environment.